MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

AN APPROACH TO CO-CHANNEL TALKER INTERFERENCE SUPPRESSION USING A SINUSOIDAL MODEL FOR SPEECH

R.G. DANISEWICZ
T.F. QUATIERI

Group 24

TECHNICAL REPORT 794

5 FEBRUARY 1988

LEXINGTON MASSACHUSETTS

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the Department of the Air Force under Contract F19628-85-C-0002.

The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Government.

The ESD Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Hugh L. Southall

Hugh L. Southall, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

ABSTRACT

A new technique for co-channel talker interference suppression has been developed, and applied effectively in situations where the speech waveforms from both the desired talker and the interfering talker are vocalic. The technique combines a minimum mean-squared error estimation procedure with a sinusoidal analysis/synthesis model of speech as a sum of sinusoids with time-varying amplitudes, frequencies, and phases. For the received waveform, which is the additive combination of two speech signals on a single-channel, least-squared error estimates of the sinusoidal model parameters for each of the two speech signals are made. A synthesizer based on the sinusoidal model is used to reconstruct the speech of the desired talker.

Initial studies of the feasibility of the new technique have examined the level of interference suppression attained when the least-squared error estimation was performed with *a priori* knowledge of (1) all the sine-wave frequencies, and (2) the fundamental frequency of each speech waveform prior to summation. In both cases, the sine-wave amplitudes and phases were unknown. When the frequencies of both waveforms were obtained by peak-picking of the individual short-time Fourier transforms prior to summation, the least-squared error strategy yielded good suppression of the interfering speech and enhancement of the target speech over a wide range (9 to −16 dB) of target-to-interferer ratios. When the individual fundamental frequency contours were provided, the enhancement was only slightly degraded. For both cases, the performance was significantly improved by a multi-frame interpolation technique which predicts the time evolution of the sinusoidal parameters across frames where the frequencies of the two waveforms are closely spaced and, therefore, difficult to track.

Finally, the least-squared-error approach was tested with no *a priori* information provided on either of the two waveforms. The least-squared error criterion was extended to estimate both fundamental frequency contours from the summed waveform, and then applied further to estimate the remaining sinusoidal parameters. This technique was demonstrated to provide useful interference suppression when the summed vocalic speech waveforms have equal intensities and smooth, nonintersecting fundamental frequency contours.

The results obtained, though limited in their scope, provide evidence that the combination of the sinusoidal analysis/synthesis model with effective parameter estimation techniques offers a promising approach to the currently-unsolved problem of co-channel talker interference suppression over a range of conditions. Promising areas for further investigation are identified.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

AN APPROACH TO CO-CHANNEL TALKER INTERFERENCE SUPPRESSION USING A SINUSOIDAL MODEL FOR SPEECH

1. INTRODUCTION

In a number of important applications, it is desirable to suppress an interfering waveform which degrades a desired signal. When the desired signal and the interfering signal are additively combined speech waveforms, the goal is to enhance the intelligibility of a target speaker. This problem is often referred to as co-channel talker interference suppression. The interfering speech may have been introduced in a microphone environment or may have resulted from cross talk in a neighboring communications channel.[1]

This report describes a new approach to co-channel talker interference suppression based on a sinusoidal representation of speech.[2,3] The technique fits a sinusoidal model to additive vocalic speech segments such that the least mean squared error between the model and the combined waveforms is obtained. Enhancement is achieved by synthesizing a waveform from the sine waves attributed to the desired speaker. Least squares estimation is applied to obtain sine-wave amplitudes and phases of both talkers, based on either *a priori* sine-wave frequencies or *a priori* fundamental frequency contours. When the frequencies of the two waveforms are closely spaced, the least squares approach can have difficulty in tracking the sine-wave parameters. In these cases, the performance is significantly improved by an interpolation technique which predicts the time evolution of the sinusoidal parameters across multiple analysis frames. The least-squared error approach is also extended to estimate fundamental frequency contours of both speakers from the summed waveform, and is applied further to estimate the remaining sinusoidal parameters.

Numerous other methods have been proposed for vocalic speech separation. One approach relies on the short time Fourier transform (STFT) of the target speech having its energy focused in regions about multiples of the target speaker's fundamental frequency. Intelligibility improvements are attempted by comb filtering the interfering speaker's harmonics from the sum.[4,5] A disadvantage of comb filtering stems from the short duration over which speech remains stationary with respect to a periodicity assumption. The duration of the comb filter's impulse response must be made correspondingly short. This constraint prevents the separation of closely spaced harmonics. One method which explicitly attempts to resolve closely spaced harmonics was introduced by Parsons[6,7] who exploited the shape of the Fourier transform of the time-domain window used in computing the STFT. Another class of methods, harmonic magnitude suppression (HMS), was introduced by Hanson and Wong[1] and further developed by Naylor and Boll,[8,9] and Childers and Lee.[10] In this approach, the short time Fourier transform magnitude (STFTM) of the summed speech is sampled at the harmonics of the interferer which is assumed to be much larger (e.g., 6 to 16 dB larger) than the target speech. These samples are used to obtain an STFTM estimate of the interfering speech which is then subtracted from the STFTM of the sum to yield an estimate of the target spectral magnitude. The phase from the original summed waveform is used to supply the phase for the estimate of the target utterance.

In addition to a new framework for the talker interference suppression problem, the sinusoidal approach of this report differs from other methods in three important ways. First, the sine-wave based method allows for closely spaced harmonics by modeling the linear dependence of the STFT on sine-wave parameters for each speaker; features in the STFT are not relied on in detecting the presence of closely spaced harmonics.[6,7] Second, the sine-wave phases, and hence the phase of the STFT of each speech waveform, is explicitly estimated. And finally, when parameter separation is difficult in regions of closely-spaced harmonics, the time evolution of model parameters is exploited.

The outline of this report is as follows. Section 2 reviews speech analysis/synthesis based on the sinusoidal model and gives the extension to the two-speaker case. Two candidate methods for talker separation with this system, *peak-picking* and *frequency-sampling*, are described. The problem of using these systems for talker separation with closely-spaced frequencies motivates the new least squares sine-wave parameter estimation methods described in Sections 4 and 5. In Section 4, knowledge of the two underlying pitch periods, or more generally the sinusoidal frequencies, is assumed. Under this condition, an equivalency is demonstrated between the least squares solution and a simple frequency-domain solution which uses the linear dependence of the STFT on sine-wave parameters. A method for estimating the pitch contours from the summed speech waveforms is discussed in Section 5. The focus of Section 6 is the multi-frame interpolation strategy for estimating the amplitudes and phases of sine waves that are not resolved by the least squares estimation method. Section 7 gives the results of informal listening tests that illustrate performance. Finally, Section 8 makes suggestions for further research.

2. THE TALKER INTERFERENCE SUPPRESSION PROBLEM IN THE CONTEXT OF THE SINE-WAVE MODEL

In this section, the sinusoidal model of speech is extended to the two-speaker case. Based on this model, a speech analysis-synthesis system is described and two methods of interference suppression are proposed. The frequency-domain properties of these approaches are discussed and the problem of separating model parameters for closely spaced frequencies is described. The section begins with a review of an analysis-synthesis system for the single-speaker case.

2.1 Speech Analysis-Synthesis Based on a Sinusoidal Model

According to the speech production mechanism, speech can be modeled as the output of a slowly-varying vocal tract filter with a quasi (i.e., "almost") periodic excitation input during voiced speech and a noise-like excitation during unvoiced speech.[1] Under this condition, the speech waveform can be represented by a sum of sine waves with time-varying amplitudes, frequencies, and phases.[2,3]

$$x(n) = \sum_{k=1}^{M} a_k(n) \cos [\theta_k(n)] \tag{2.1}$$

where the amplitudes and phases, are denoted by $a_k(n)$ and $\theta_k(n)$, respectively, and the time-varying frequency of each sine wave is given by the derivative of the phase and will be denoted by $\omega_k(n) = \theta'_k(n)$. If the excitation is periodic, with a slowly-varying period, then the frequencies can be represented by multiples of a slowly-varying fundamental frequency, $\omega_o(n)$. This harmonic model and the more general model (2.1) will be used in the various systems throughout this report. For the purpose of designing an analysis-synthesis system based on the model (2.1), we simplify the phase function $\theta_k(n)$ by assuming a fixed frequency ω_k and fixed amplitude a_k in a time interval over which waveform analysis will take place (typically 20 to 30 ms). Under this condition, the model in (2.1) is given by

$$x(n) = \sum_{k=1}^{M} a_k \cos [\omega_k n + \phi_k] \tag{2.2a}$$

where

$$\theta_k(n) = \omega_k n + \phi_k \tag{2.2b}$$

and where the phase value ϕ_k is the phase offset measured relative to the origin of the analysis frame (i.e., n = 0).

An overview of an analysis-synthesis system based on this sinusoidal model (2.1) and (2.2) is depicted in Figure 2-1 (References 2 and 3). A short-time Fourier transform (STFT) is computed every 10 to 20 ms over a 20 to 30 ms analysis window using the discrete Fourier transform (DFT). The frequencies ω_k are estimated by picking the peaks of the uniformly spaced samples of

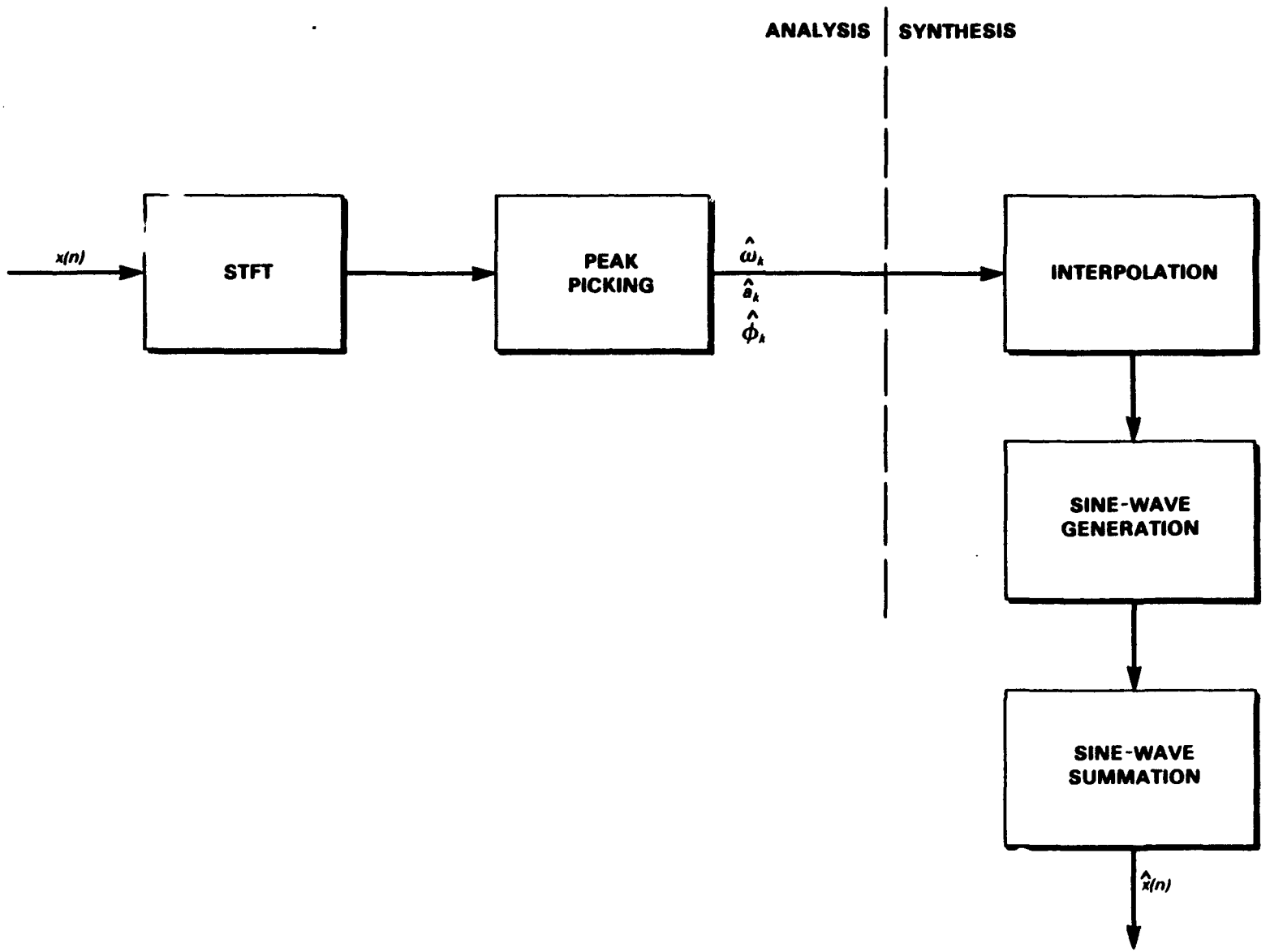


Figure 2-1. Sinusoidal analysis-synthesis.

89335-1

the short-time Fourier transform magnitude (STFTM). Alternatively, the frequencies can be estimated via an estimate of the fundamental frequency. The sine-wave amplitudes a_k and phases ϕ_k for each analysis frame are then given by the amplitude and phase of the STFT at the measured frequencies.

The first step in the synthesis requires association of the frequencies measured on one frame with those obtained on a successive frame. This is accomplished with a nearest-neighbor matching algorithm which incorporates a birth/death process of the component sine waves.[2] Amplitude and phase parameters are then interpolated across frame boundaries at these matched frequency sets. Since the amplitudes are slowly varying, it suffices to interpolate them linearly. In interpolating the phase, since the phase is measured modulo 2π, phase unwrapping must be performed. In addition, since the phase is the integral of the instantaneous frequency, the interpolation must yield a phase which is consistent with the frequencies measured at each frame boundary. To solve this problem, a cubic polynomial is used for the interpolation function. The solution requires constraining the cubic function and its derivative to equal the measured phases and frequencies, respectively, at the frame boundaries, and to satisfy a smoothness constraint across each frame. When the sine-wave frequencies are measured via peak picking, the waveform estimate, $\hat{x}(n)$, obtained by summing the amplitude-modulated sine waves, is perceptually nearly indistinguishable from the original.

2.2 The Two-Speaker Case

The sinusoidal speech model for the single-speaker case is easily generalized to the two-speaker case. A speech waveform generated by two simultaneous talkers can be represented by a sum of two sets of sine waves each with time-varying amplitudes, frequencies, and phases

$$x(n) = x_a(n) + x_b(n) \quad . \tag{2.3}$$

$$x_a(n) = \sum_{k=1}^{M_a} a_k(n) \cos\left[\theta_{a,k}(n)\right]$$

$$x_b(n) = \sum_{k=1}^{M_b} b_k(n) \cos\left[\theta_{b,k}(n)\right]$$

where the sequences, $x_a(n)$ and $x_b(n)$ denote the speech of speaker A and the speech of speaker B, respectively. The amplitudes and phases associated with speaker A are denoted by $a_k(n)$ and $\theta_{a,k}(n)$ and the frequencies are given by $\omega_{a,k}(n) = \theta'_{a,k}(n)$. A similar parameter set is associated with speaker B. If the excitation is periodic, a two-speaker harmonic model can be used where the frequencies associated with speaker A and speaker B are multiples of two underlying fundamental frequencies, $\omega_a(n)$ and $\omega_b(n)$, respectively. In the steady-state case where the vocal cords and vocal tract characteristics are assumed fixed over the analysis time interval, we can write the model of (2.3) as

$$x(n) = x_a(n) + x_b(n) \quad . \tag{2.4}$$

$$x_a(n) = \sum_{k=1}^{M_a} a_k \cos(\omega_{a,k} n + \phi_{a,k})$$

$$x_b(n) = \sum_{k=1}^{M_b} b_k \cos(\omega_{b,k} n + \phi_{b,k})$$

which is a useful model on which to base sine-wave analysis.

Using the model (2.3) and (2.4), it is possible, as in the single speaker case, to reconstruct the two-speaker waveform with the analysis-synthesis system illustrated in Figure 2-1. In order to obtain an accurate representation of the waveform, the number of sine waves in the underlying model is chosen to account for the presence of two speakers. The presence of two speakers also requires that the analysis window length be chosen to resolve frequencies more closely spaced than in the single-speaker case. Due to the requirement of time resolution, however, the analysis window length was chosen to give adequate frequency resolution for the lower-pitch speaker. The reconstruction yields synthetic speech that is again nearly indistinguishable from the original summed waveform.[2,3]

The ability to recover the summed waveform via the analysis-synthesis system of Figure 2-1 suggests the method of *peak picking* in Figure 2-2 for recovering a desired waveform $x_b(n)$ which is of lower intensity than an interfering background talker $x_a(n)$. The largest peaks of the summed spectra (the number of peaks is equal to or less than the number required to represent a single waveform) are chosen and are used to reconstruct the larger of the two waveforms. This waveform estimate is then subtracted from the combined waveform to form an estimate of the lower signal. The largest peaks of the summed spectra, however, do not necessarily represent the peaks of the spectra of the larger waveform; that is, they will in general contain information about both waveforms. The parameters which form the basis for the reconstruction of the summed waveforms do not necessarily form the basis for reconstructing the individual speech waveforms.

An example is illustrated in Figure 2-3 which shows the sine-wave frequency tracks derived from combined voiced segments of male and female speech. Since the female speaker is 18 dB above the male speaker, we expect the larger peaks to represent the peaks of the female speaker. In fact, as shown in Figure 2-3(c), the frequency tracks derived from the largest peaks appear to correspond to primarily the tracks of the larger speaker. The waveform reconstructed using these frequency tracks and corresponding amplitudes and phases of the STFT, however, manifests the lower speaker almost as clearly as in the original summed waveform. The attempted recovery of the larger speaker with only 25, 10, and even 5 of the largest peaks still allows the lower speaker to be heard in the synthesis. Performing the subtraction indicated in Figure 2-2 results in a "garbled" reconstruction with no apparent enhancement of $x_b(n)$. These results held for summed

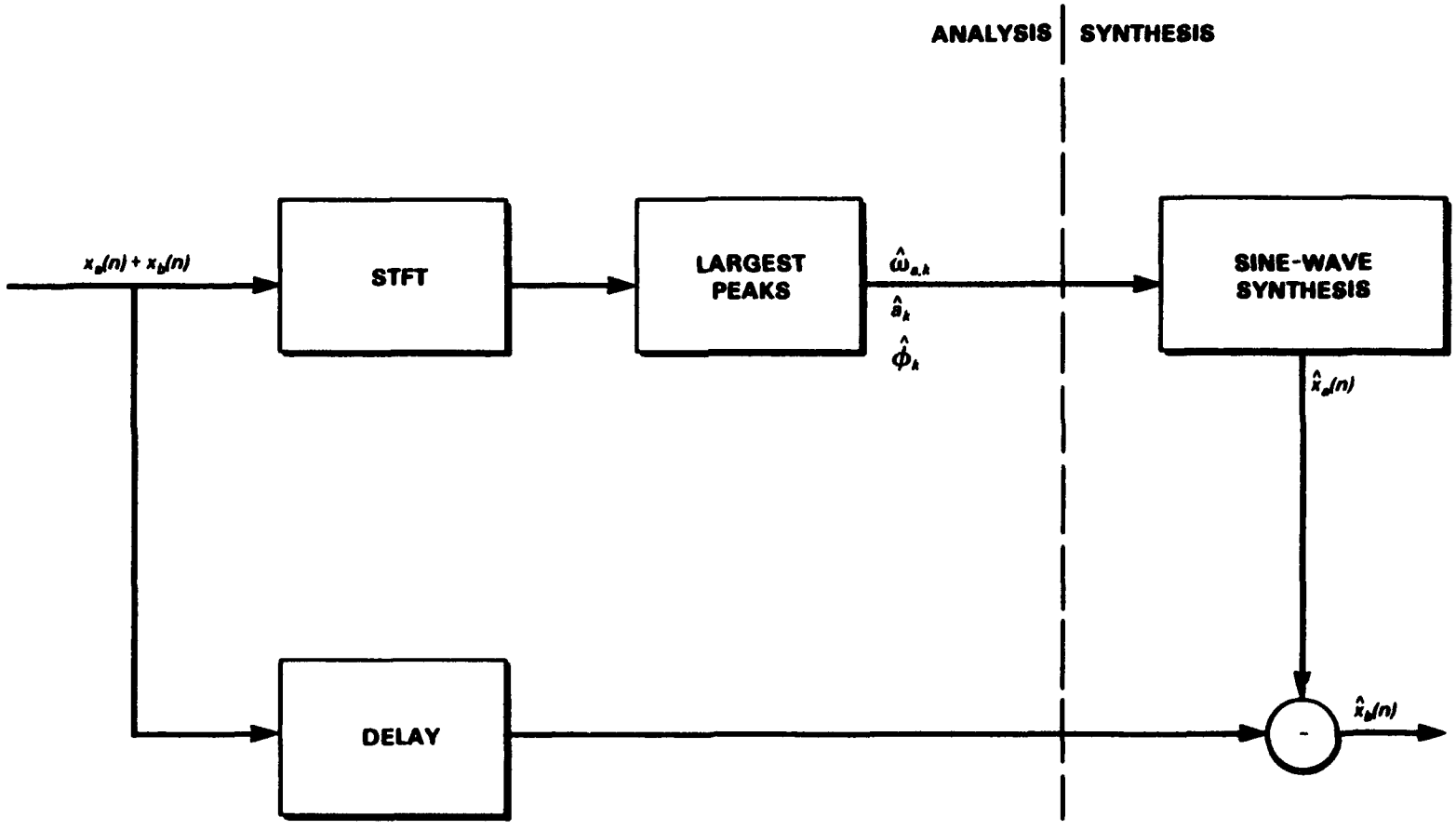


Figure 2-2. Sinusoidal reconstruction of speech from the summed waveform.

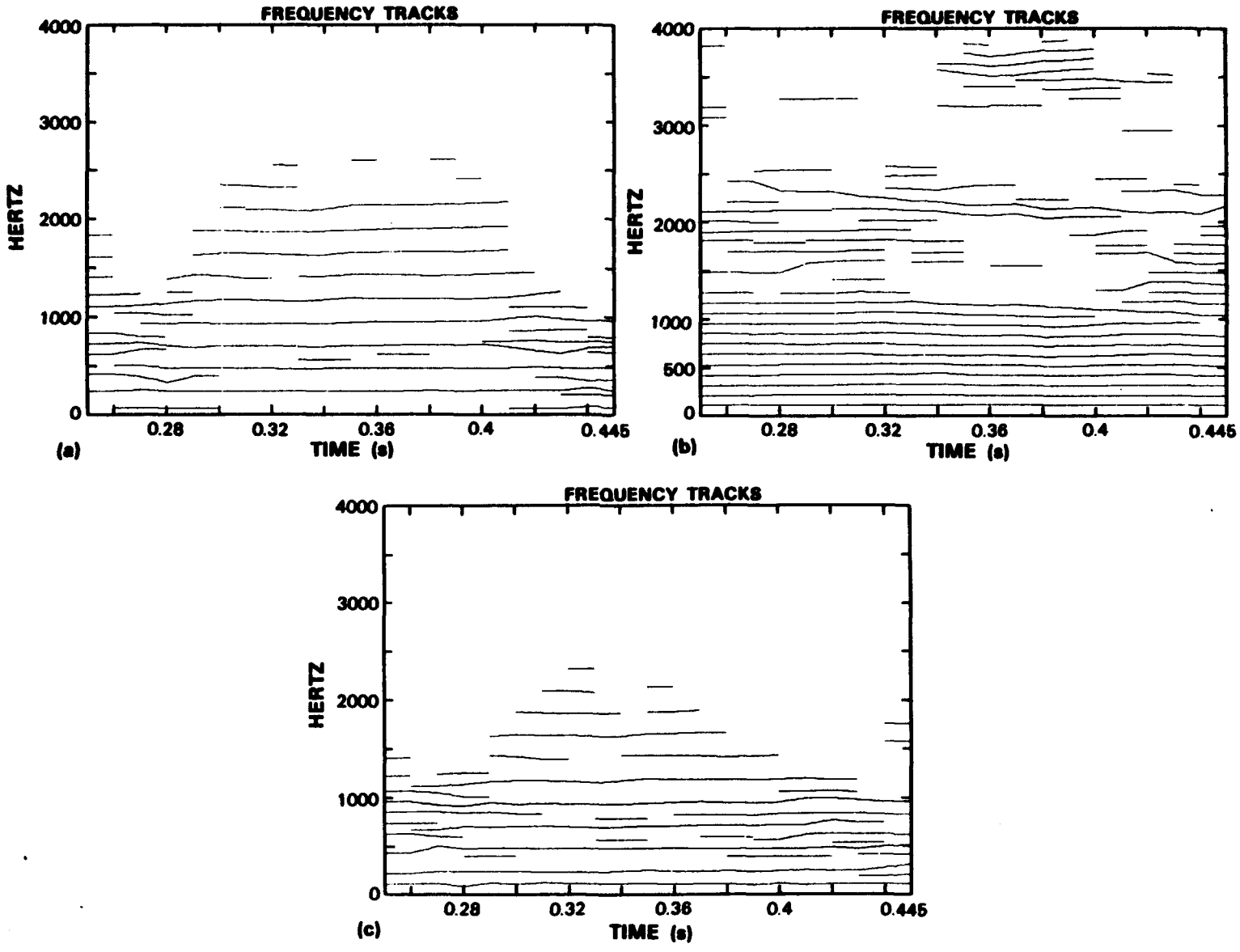


Figure 2-3. Frequency tracks of additively combined male and female speech (a) tracks of female speaker,(b) tracks of male speaker,and (c) tracks of combined speakers.

all-voiced passages, as well as summed voiced and unvoiced passages. One problem with this technique, described in the following section, is that closely spaced frequencies associated with different speakers may be seen as one peak by the peak-picking process. When the analysis window length is made very large, e.g., 50 ms, thus improving frequency resolution, the lower speaker is further suppressed in the reconstruction of $\hat{x}_a(n)$. Even in this case, however, inadequate time resolution, as well as closely-spaced frequencies, prevent the enhancement of $x_b(n)$ by the interference suppression system of Figure 2-2.

Another approach to separating two summed passages via sine-wave analysis-synthesis is illustrated in Figure 2-4. Knowledge of the sine-wave frequencies of each of the two speakers is assumed; the frequency sets are obtained by peak-picking individual STFTMs and then are used to sample the summed STFT to obtain amplitude and phase estimates for the sine-wave representation of each waveform. An estimate of the desired lower waveform, $\hat{x}_b(n)$, could then be directly reconstructed or, as illustrated in Figure 2-4, an estimate, $\tilde{x}_b(n)$, can be obtained by subtracting the reconstructed larger waveform from the summed waveform. (The linearity of the STFT operator makes these two estimates essentially equivalent.) We refer to this approach as *frequency sampling* since the summed STFT is sampled at the known frequencies. Alternatively, the frequency sets might be obtained by estimating a fundamental frequency for each speaker. This method is akin to comb filtering which extracts a waveform by processing the sum with a filter derived by placing its resonances about multiples of an assumed fundamental frequency.[4,5] Although these methods use more accurate frequency estimates than from peak-picking the summed STFTM, the accuracy of the corresponding amplitudes and phases is limited, as before, by the tendency of frequencies of the two waveforms to often be closely spaced. As we will see in Section 6, enhancement obtained by the method of frequency sampling, therefore, is small, in spite of the large *a priori* information required by this method.

2.3 The Problem of Closely Spaced Frequencies

In the last section we saw that although the summed waveform $x(n) = x_a(n) + x_b(n)$ is well represented by peaks in the STFT of x(n), the sine-wave amplitudes and phases of the individual waveforms are not easily extracted from these values. In this section we investigate the problem of extracting the sine-wave amplitudes and phases of $x_a(n)$ and $x_b(n)$ from the STFT of x(n).

Let $s_p(n)$ represent the pth windowed speech segment extracted from a time-shifted version of the sum of two sequences

$$s_p(n) = w(n)[x_a(n + pL) + x_b(n + pL)] \quad ; \quad \frac{-(N-1)}{2} < n < \frac{N-1}{2} \tag{2.5}$$

where L is the time shift between segments and where w(n) is nonzero over the interval $-(N-1)/2 < n < (N-1)/2$, and, in this study, is given by the Hanning window.[12] The short time Fourier transform (STFT) of the summed waveforms, $S_p(\omega)$, is given by

$$S_p(\omega) = \sum_{n=-(N-1)/2}^{(N-1)/2} s_p(n)\, e^{-j\omega n} \tag{2.6}$$

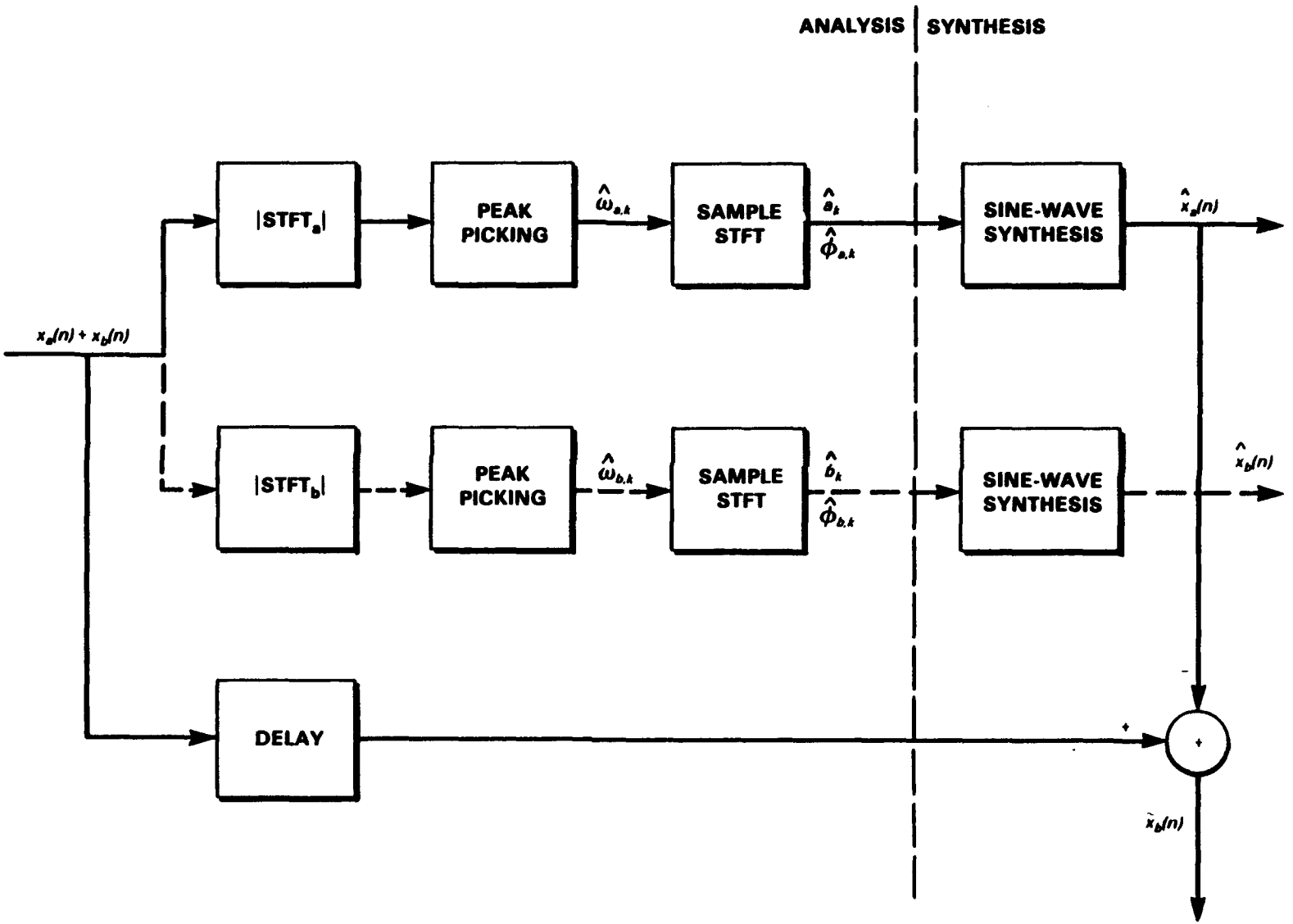


Figure 2-4. Reconstruction by frequency sampling.

89336-6

which, in practice, is computed at uniformly spaced samples with the fast Fourier transform (FFT).[12] By substituting (2.4) and (2.5) into (2.6), Equation (2.6) can be written as a summation of scaled and shifted versions of the transform of the analysis window.

$$S_p(\omega) = \sum_{k=1}^{M_a} \frac{1}{2} a_k \exp(j\phi_{a,k}) W(\omega - \omega_{a,k}) + \sum_{k=-1}^{M_a} \frac{1}{2} a_k \exp(-j\phi_{a,k}) W(\omega + \omega_{a,k}) \qquad (2.7)$$

$$+ \sum_{k=1}^{M_b} \frac{1}{2} b_k \exp(j\phi_{b,k}) W(\omega - \omega_{b,k}) + \sum_{k=-1}^{M_b} \frac{1}{2} b_k \exp(-j\phi_{b,k}) W(\omega + \omega_{b,k})$$

where $W(\omega)$ denotes the Fourier transform of the analysis window w(n) and where for simplicity we assume that the time shift of the analysis frame in (2.5) is zero.

The success of extracting sine-wave parameters by peak-picking, described in the previous Section 2.2, depends on the properties of $W(\omega)$, the Fourier transform of the analysis window. The effective bandwidth of $W(\omega)$ is inversely proportional to N, the duration of the analysis window. Longer window lengths give rise to narrower spectral main lobes.[12] If the spacing between the shifted versions of $W(\omega)$ in (2.7) is such that the main lobes do not overlap, a reasonable strategy for extracting the model frequencies and performing the separation is the method of *peak-picking*. For the case of summed speech waveforms, however, this constraint is not often met since the analysis window cannot be made arbitrarily large. Even when the frequencies are known *a priori*, i.e., the method of *frequency sampling* is used, when the frequencies are closely spaced, accurate estimates of the sine-wave amplitudes and phases are generally not obtained.

Figure 2-5 illustrates an example where the frequencies are spaced closely enough to prevent accurate separation by the above methods. Figures 2-5(a) and 2-5(b) depict the STFTM of two steady-state vowels over a 25 ms interval. The vowels have roughly equal intensity and belong to two speakers who have dissimilar fundamental frequencies. The STFTM of the summed waveforms appears in Figure 2-5(c). A subset of the main lobes of the Fourier transform of the analysis windows overlap and add such that they merge to form a single composite lobe (since the addition is complex, lobes may destructively interfere as well). When the peak-picking strategy is applied to the STFTM of the summed speech waveform, the process may allot a single frequency to represent these composite structures. For this reason, the frequency sampling strategy will also have difficulty in recovering the individual sine-wave amplitude and phase parameters.

One approach to extracting the underlying amplitude and phase of the STFT of $x_a(n)$ and $x_b(n)$ is to detect the presence of overlap and then use the structure of the analysis window in the frequency domain to help in the separation.[6,7] Figure 2-5 shows that "features" in the STFTM of x(n) are not, however, reliable in detecting the presence of a single composite lobe formed by two overlapping lobes. Unique characteristics in the phase of the STFT (depicted by dotted lines in Figure 2-5) of overlapping lobes are also difficult to determine. For example, the two largest lobes in the summed spectra are characterized by both magnitude symmetry and a flat phase characteristic which characterizes either speaker A or speaker B. Thus any technique for separation relying on such features will be prone to error.

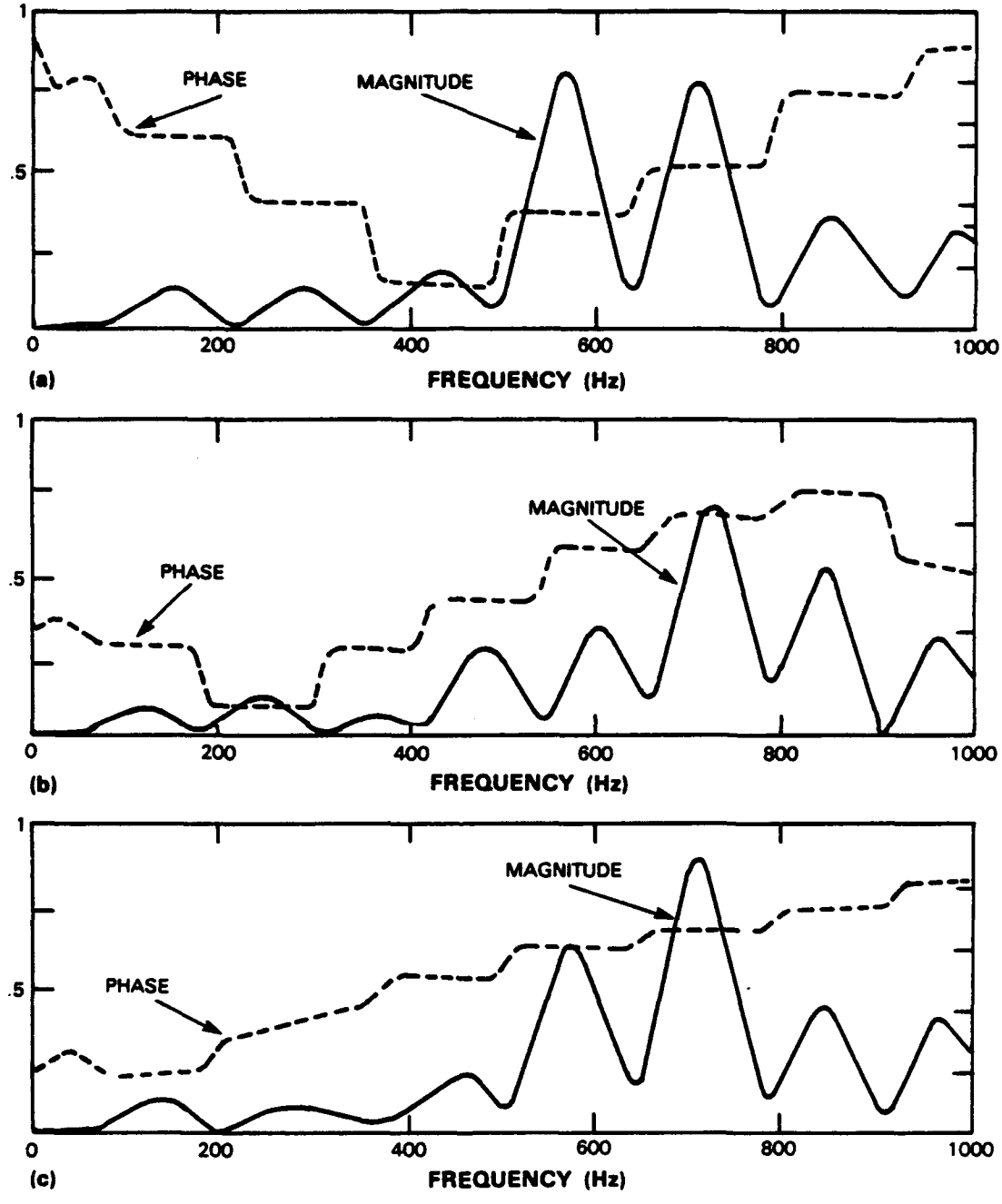


Figure 2-5. Properties of the STFT of $x(n) = x_a(n) + x_b(n)$.
(a) STFT magnitude and phase of $x_a(n)$.
(b) STFT magnitude and phase of $x_b(n)$.
(c) STFT magnitude and phase of $x(n) = x_a(n) + x_b(n)$.

2.4 The Least Squares Approach

The discussion of the previous section suggests that the linear combination of the shifted and scaled Fourier transforms of the analysis window in (2.7) must be explicitly accounted for in achieving separation. The (complex) scale factor applied to each such transform corresponds to the desired sine-wave amplitude and phase, and the location of each transform is the desired sine-wave frequency. Parameter estimation is difficult, however, due to the nonlinear dependence of the sine-wave representation on phase and frequency.

The approach to separation in this report first assumes *a priori* frequency knowledge, and performs a least squares fit to the summed waveform with respect to the unknown sine-wave amplitude and phase parameters. In the next section we show that this solution is equivalent to solving for the sine-wave amplitudes and phases via the linear relationships suggested by (2.7). Figure 2-6 puts this least squares approach in perspective with other strategies we have discussed and which we will further compare in the sequel. In contrast to the frequency sampling method which samples the STFT at the known frequencies, in one implementation of the least squares method, the frequency sampling solution is used as an initial guess and iterated upon to form more accurate estimates. Figure 2-6 shows that the frequencies themselves may also be estimated via a least squares formulation. In Section 4, this estimation problem will be simplified by constraining the frequencies to be harmonically related.

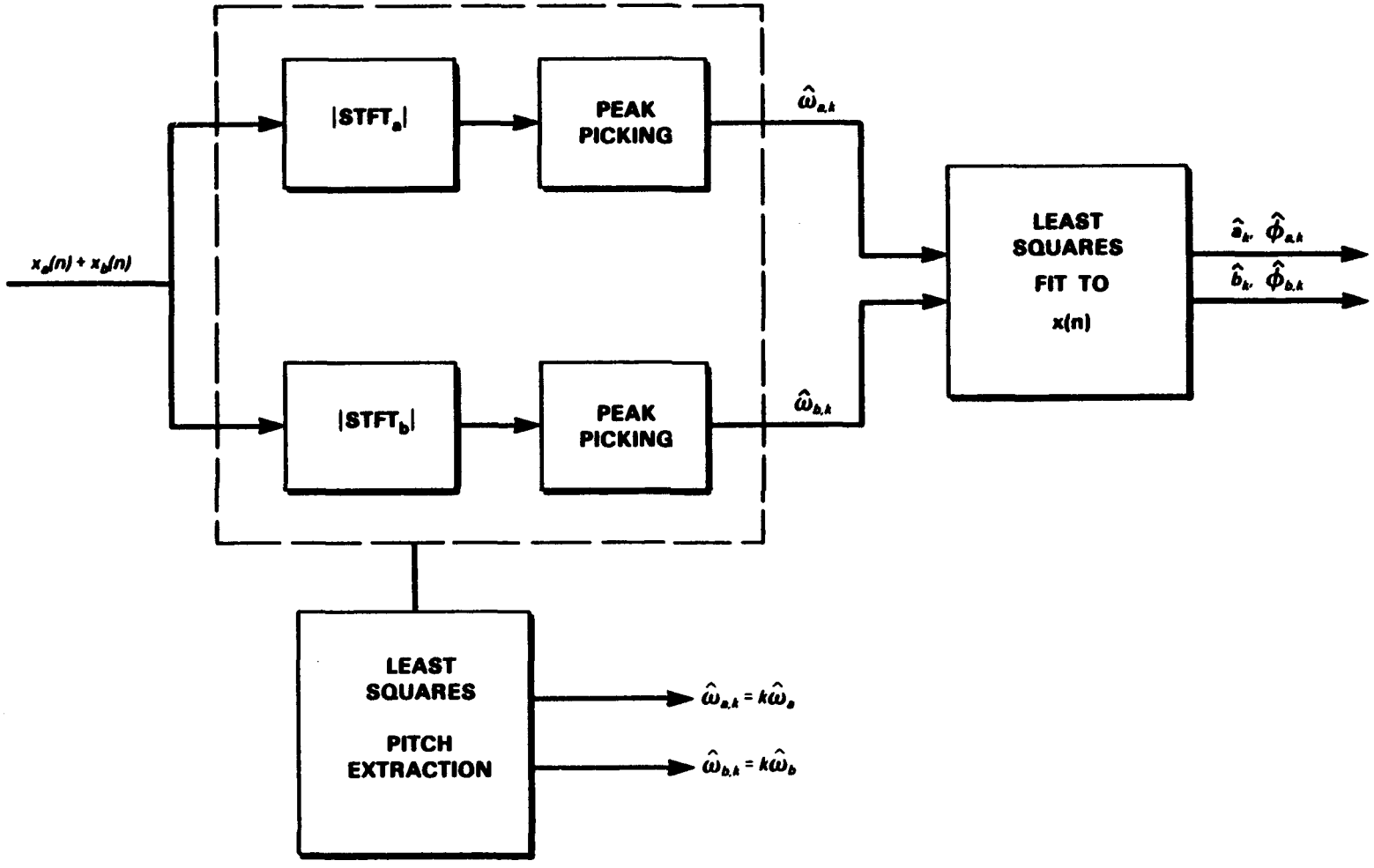


Figure 2-6. The least squares approach.

89335-8

3. THE LEAST SQUARES SOLUTION

In this section, we transform the nonlinear problem of forming a least squares solution for the sine-wave amplitudes, phases, and frequencies into a linear problem. We accomplish this by assuming the sine-wave frequencies are known *a priori*, and by solving for the real and imaginary components of the quadrature representation of the sine waves, rather than solving for the sine-wave amplitudes and phases. The previous section suggests that these parameters can be obtained by exploiting the linear dependence of the STFT on scaled and shifted versions of the Fourier transform of the analysis window. We begin this section with a solution based on this observation, and then show that the parameters derived by this approach represent the sine-wave parameters chosen by forming a least squares fit to the summed speech waveforms.

3.1 Solving for the Sine-Wave Parameters

Figure 3-1 illustrates how the main lobes of two shifted versions of the Fourier transform of the analysis window, $W(\omega)$, typically overlap when they are centered at two closely spaced frequencies ω_1 and ω_2, corresponding to speaker A and speaker B, respectively, each consisting of a single frequency. Figure 3-1 suggests a strategy for separation by solving the following linear equations

$$\begin{bmatrix} 1 & W(\Delta\omega) \\ W(\Delta\omega) & 1 \end{bmatrix} \begin{bmatrix} S_a(\omega_1) \\ S_b(\omega_2) \end{bmatrix} = \begin{bmatrix} S(\omega_1) \\ S(\omega_2) \end{bmatrix} \tag{3.1}$$

where $S_a(\omega_1)$ and $S_b(\omega_2)$ denote the samples of the STFTs at known frequencies ω_1 and ω_2, and $\Delta\omega$ is the distance in frequency between them. The amplitudes and phases of $S_a(\omega_1)$ and $S_b(\omega_2)$ represent the unknown parameters of the two underlying sine waves. The STFT of the sum is denoted by $S(\omega)$ [for simplicity the subscript "p" in (2.7) has been removed]. The Fourier transform of the analysis window is denoted by $W(\omega)$ with normalization $W(0) = 1$. Since the window transform is real, the matrix in the left side of (3.1) is real; however, the STFT of the waveform is complex, so that the complex solution to (3.1) can be obtained by solving separately the real and imaginary parts of the matrix equation. Equation (3.1) is not exact since the contribution from the Fourier transforms of the analysis window centered at $-\omega_1$ and $-\omega_2$ has not been included (in practice, the signal to be transformed is real and so both positive and negative frequency contributions will exist). For simplicity, we assume this contribution is negligible.

Since from (2.7) the STFT of a sum of windowed sinusoids is a sum of shifted and scaled versions of $W(\omega)$, the two-lobe case of Figure 3-1 can be simply extended to the case where there are M overlapping lobes of the form in Figure 3-1. Specifically, a relation can be written which reflects the linear dependence of the STFT on all M lobes (see Appendix A).

$$\mathbf{H}\underline{\alpha} = 2\mathrm{Re}\,[\underline{S}(\underline{\omega})] \tag{3.2a}$$

$$\mathbf{H}\underline{\beta} = -2\mathrm{Im}\,[\underline{S}(\underline{\omega})] \tag{3.2b}$$

where $\underline{S}(\underline{\omega})$, consisting of STFT samples, is a vector function of the sinusoidal frequency vector $\underline{\omega}$ given by

$$\underline{\omega} = (\omega_1, \omega_2, \omega_3, \ldots\ldots, \omega_M)^T \quad ; \quad \omega_1 < \omega_2 < \omega_3 < \ldots \omega_M \tag{3.3}$$

consisting of frequencies from both speaker A and speaker B and where,

$$\mathbf{H} = \begin{bmatrix} W(0) & W(\omega_1 - \omega_2) & W(\omega_1 - \omega_3) & & W(\omega_1 - \omega_M) \\ W(\omega_2 - \omega_1) & W(0) & W(\omega_2 - \omega_3) & & W(\omega_2 - \omega_M) \\ W(\omega_3 - \omega_1) & W(\omega_3 - \omega_2) & W(0) & \cdots & W(\omega_3 - \omega_M) \\ W(\omega_4 - \omega_1) & W(\omega_4 - \omega_2) & W(\omega_4 - \omega_3) & & W(\omega_4 - \omega_M) \\ & & \vdots & & \\ W(\omega_M - \omega_1) & W(\omega_M - \omega_2) & W(\omega_M - \omega_3) & & W(0) \end{bmatrix} \tag{3.4}$$

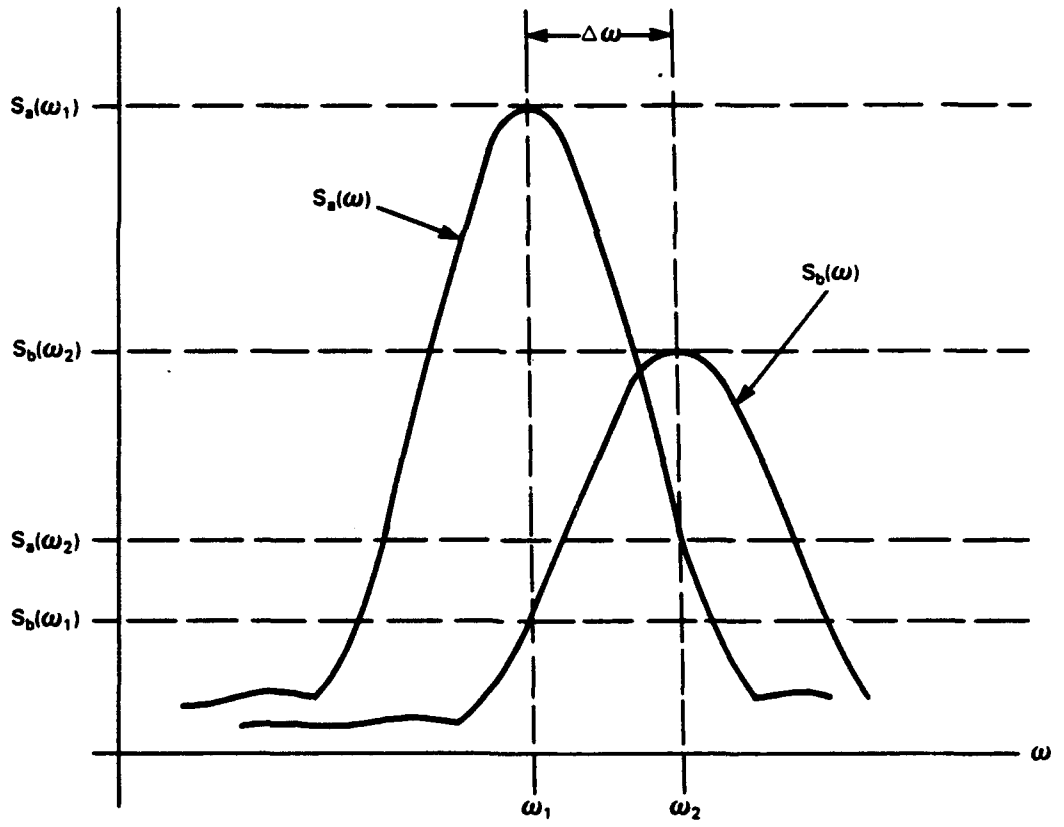


Figure 3-1. Two overlapping main lobes of shifted and scaled versions of W(ω).

The vectors $\underline{\alpha}$ and $\underline{\beta}$ consist of estimates of the unknown parameters of (2.4) but in quadrature form

$$\hat{x}(n) = \sum_{k=0}^{M} \alpha_k \cos(\omega_k n) + \sum_{k=0}^{M} \beta_k \sin(\omega_k n) \tag{3.5a}$$

with

$$\alpha_k = \hat{a}_k \cos(\hat{\phi}_k) \tag{3.5b}$$

and

$$\beta_k = \hat{a}_k \sin(\hat{\phi}_k) \tag{3.5c}$$

where "^" denotes estimate and where $M = M_a + M_b$. Equation (3.5) can also be expressed in terms of polar coordinates

$$\hat{x}(n) = \sum_{k=1}^{M} c_k \cos(\omega_k n + \hat{\phi}_k) \tag{3.6}$$

$$c_k = \sqrt{\alpha_k^2 + \beta_k^2} \quad , \quad \hat{\phi}_k = \tan^{-1}(\beta_k / \alpha_k)$$

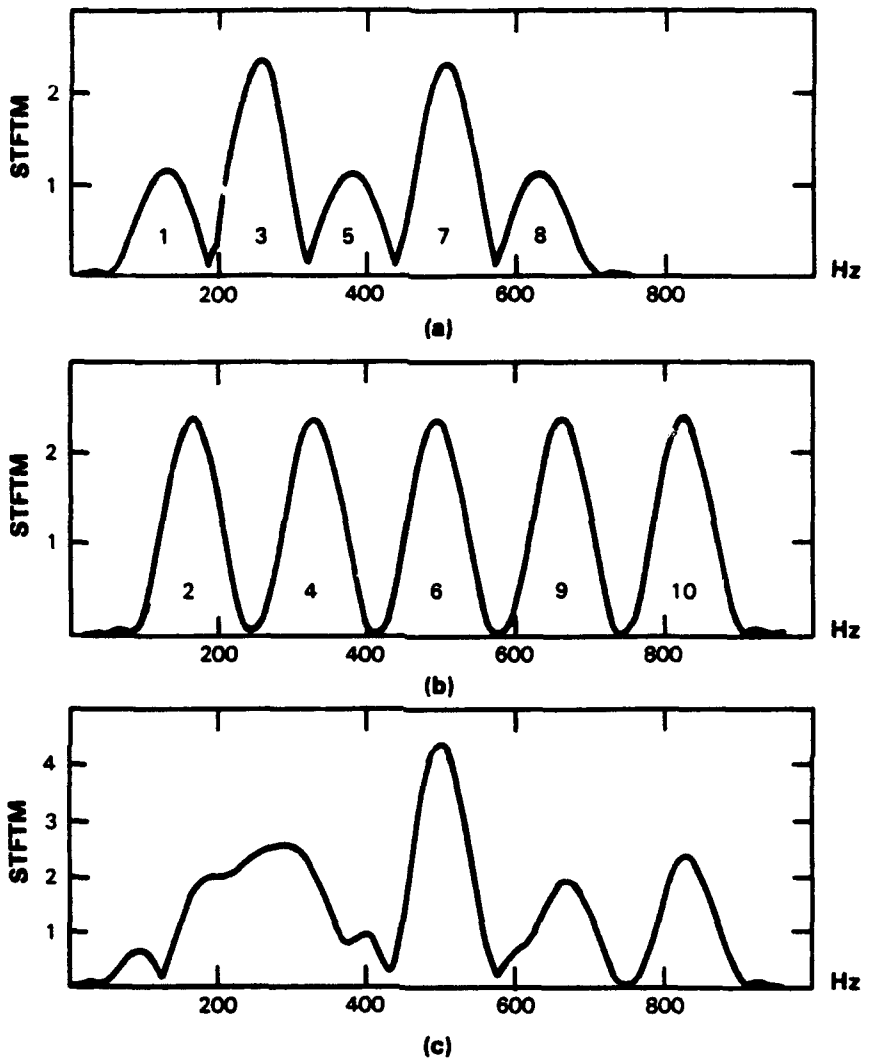
For speaker separation, Equation (3.6) can be partitioned since we assume the partitioning of the frequency vector $\underline{\omega}$ is known *a priori*

$$\hat{x}(n) = \sum_{k=1}^{M_a} \hat{a}_k \cos(\omega_{a,k} n + \hat{\phi}_{a,k}) + \sum_{k=1}^{M_b} \hat{b}_k \cos(\omega_{b,k} n + \hat{\phi}_{b,k}) \tag{3.7}$$

and thus solution to the matrix equation (3.2) yields the sine-wave amplitudes and phases of the two underlying speech components.

Figure 3-2 gives an example of the STFTM of two summed frames of vocalic speech and Figure 3-3 shows the corresponding **H** matrix. Although the **H** matrix has values that occur off of the main diagonal, these values fall off rapidly as the distance from the main diagonal increases. This property reflects the condition that overlap among the main lobes of scaled and shifted versions of the Fourier transform of the window occurs primarily between neighboring lobes of different speakers (the analysis window is assumed long enough so that main lobes of a single speaker do not overlap). Occasionally, however, the **H** matrix will have a broader diagonal arising when the speakers are low in pitch and the window lengths are short in duration.

The preceding analysis views the problem of solving for the sine-wave amplitudes and phases in the frequency domain. Alternatively, the problem can be viewed in the time domain. In Appendix B, it is shown that, for suitable window lengths, the vectors $\underline{\alpha}$ and $\underline{\beta}$ that satisfy (3.2) also approximate the vectors that minimize the weighted mean square distance between the measured speech frame, s(n), and the steady state sinusoidal model, x(n), for summed vocalic speech with the sinusoidal frequency vector $\underline{\omega}$ and unknown parameters $\underline{\alpha}$ and $\underline{\beta}$. Specifically, the following minimization is performed with respect to $\underline{\alpha}$ and $\underline{\beta}$



INDEX	1	2	3	4	5	6	7	8	9	10	
FREQUENCY	130	170	260	340	390	510	520	650	680	850	Hz

(d)

89336-11

Figure 3-2. Demonstration of two-lobe overlap: (a) STFTM of $x_a(n)$, (b) STFTM of $x_b(n)$, (c) STFTM of $x(n) = x_a(n) + x_b(n)$, and (d) sine-wave frequencies.

	1	2	3	4	5	6	7	8	9	10
1	.99	.52	.01	.00	.00	.00	.00	.00	.00	.00
2	.52	1.0	-.02	.00	.00	.00	.00	.00	.00	.00
3	.01	-.02	1.0	.01	.01	.00	.00	.00	.00	.00
4	.00	.00	.01	1.0	.34	.00	.00	.00	.00	.00
5	.00	.00	.01	.34	1.00	.00	.01	.00	.00	.00
6	.00	.00	.00	.00	.00	1.0	.96	.01	.00	.00
7	.00	.00	.00	.00	.01	.96	1.0	.01	.00	.00
8	.00	.00	.00	.00	.00	.01	.01	1.0	.70	.00
9	.00	.00	.00	.00	.00	.00	.00	.70	1.0	.00
10	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0

89335-10

Figure 3-3. H matrix for the example in Figure 3-2.

$$\min \sum_{n=-(N-1)/2}^{(N-1)/2} w(n)\,[x(n) - s(n)]^2 \tag{3.8}$$

The error weighting in the least-squared error (LSE) problem (3.8) is the analysis window that is used to obtain the STFT. Solution to x(n) in (3.8) is given by $\hat{x}(n)$ in (3.5). Thus the matrix equation in (3.2) can be arrived at by two apparently different approaches; in the frequency domain, by investigating the linear dependence of the STFT on scaled and shifted versions of the Fourier transform of the analysis window, or, in the time domain, by the waveform minimization given in (3.8). These two interpretations have analogies in the one-speaker case where least-squares minimization in the time domain leads to a solution which chooses sine-wave amplitudes and phases at peaks in the STFT.[2]

3.2 Implementation

The frequency estimates used in the solution (3.2) were obtained by peak-picking the STFTM of each separate waveform. A 4096-point FFT was found to give sufficient frequency resolution for adequate separation. The Gauss Seidel iterative method[13] was then used in solving (3.2). The algorithm is computationally inexpensive, easy to program, and exhibited stability and rapid convergence for the cases tested. Convergence of this algorithm is guaranteed for positive definite matrices, a property of the matrices in our least-squares problem.

The vector obtained by sampling the STFT at the sine-wave frequencies was used as an initial guess in the iterative algorithm, i.e., the solution from the frequency-sampling method. The iterative approach to solving (3.2) may be looked upon, therefore, as improving the initial amplitude and phase estimates obtained by sampling the summed STFT at the frequencies of each speaker.

3.3 An Example

An example is given in Figure 3-4 which shows the LSE solution of the two signals making up the summed speech, where speech waveform A is about 3 dB below speech waveform B. Both the waveform and STFTM estimates are compared with their respective originals. The figure gives the outcome of 50 iterations of the Gauss-Seidel algorithm, although in practice fewer iterations are required.

3.4 Sensitivity

As frequencies of speaker A come arbitrarily close to those of speaker B, the conditioning of the **H** matrix deteriorates to where the matrix becomes singular (see Appendix C). For these cases, solving the LSE problem does not permit separation. In detecting these cases, the spacing between neighboring frequencies is monitored. A single sinusoid is used to represent two sinusoids whose frequencies are closely spaced, e.g., less than 25 Hz apart. Closely-spaced frequencies which satisfy this criterion are then combined as single entries in the LSE Equations (3.2) and (3.3).

Figures 3-5(a) and 3-5(b) illustrate such an example where speaker B is 20 dB below speaker A. One lobe is missing in the reconstructed STFTM of each speaker. The monitoring procedure detected the presence of two frequencies which are close enough to cause ill-conditioning of the **H** matrix. These frequencies, merged as one in the LSE solution, were not used in the reconstruction. A strategy for resolving these ambiguities is proposed in Section 5.2.

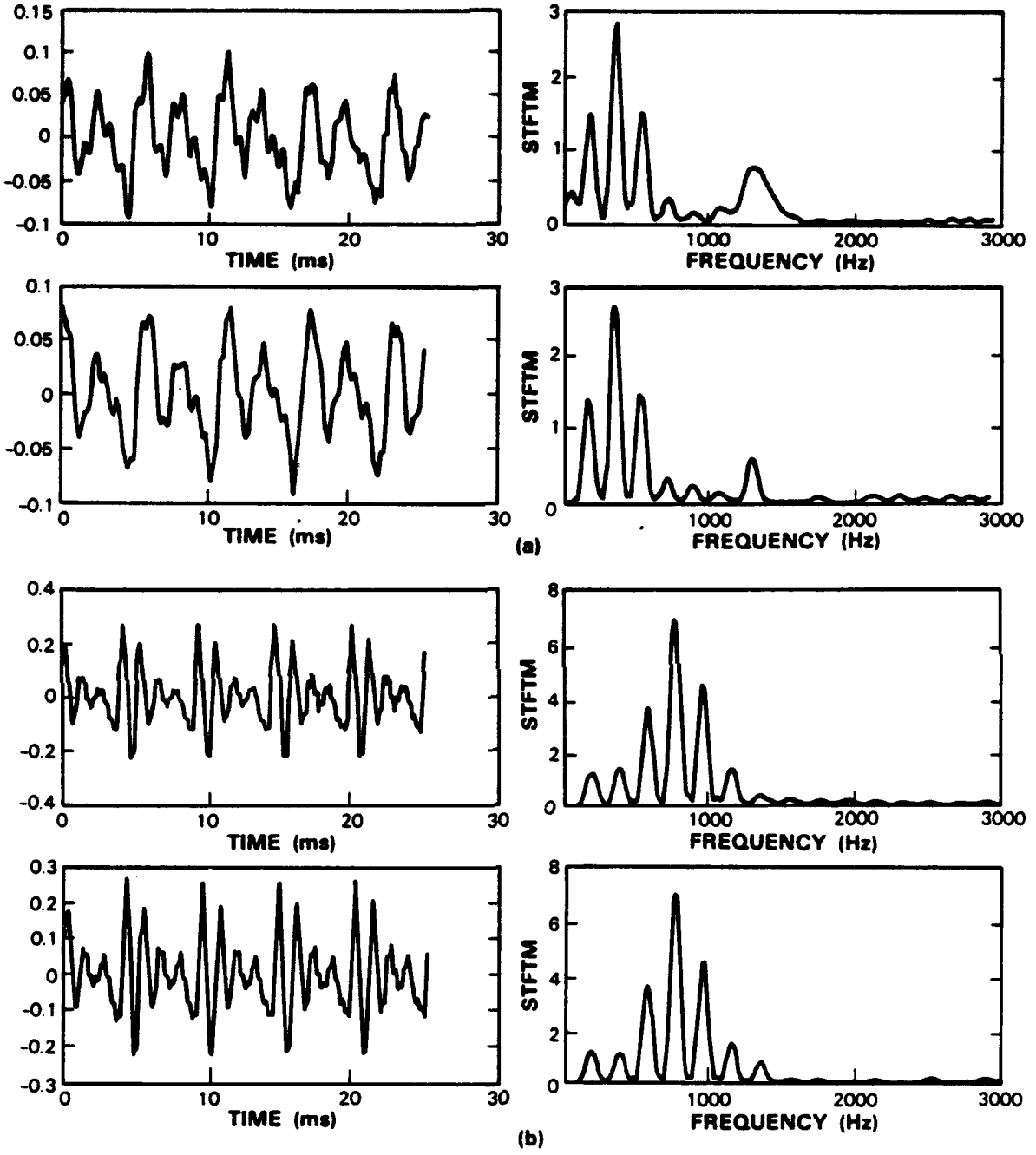


Figure 3-4. Separation of summed speech waveforms:(a) Speaker A(upper) compared to estimate of speaker A(lower)and (b) speaker B(upper) compared to estimate of speaker B(lower).

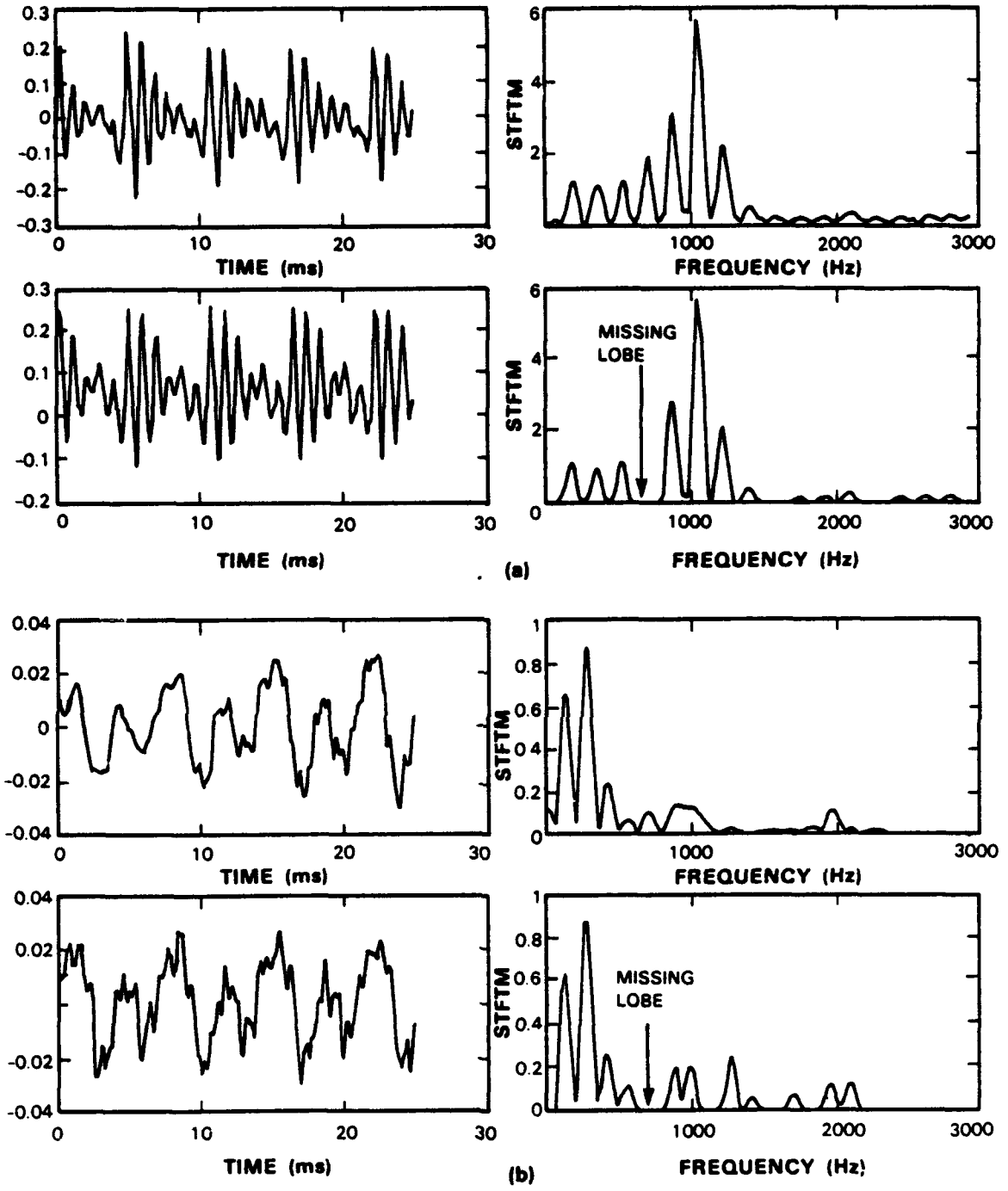


Figure 3-5. Demonstration of ill-conditioning of the H matrix: (a) speaker A(upper) compared to estimate of speaker A(lower) and (b) speaker B(upper) compared to estimate of speaker B(lower).

4. TRACKING THE FUNDAMENTAL FREQUENCY PAIR

Simultaneous estimation of sine-wave amplitudes, phases, and frequencies is a difficult nonlinear problem; the assumption of *a priori* frequency knowledge, relied on in the previous section, helped to make the problem linear. In this section, we reduce the dimensionality of the nonlinear problem by assuming frequencies for each speaker can be represented by a multiple of a fundamental frequency, ω_a for speaker A and ω_b for speaker B.

Under this harmonic assumption, since the model (2.3) and (2.4) is a nonlinear function of the fundamental frequencies, a simple closed-form solution, based on the least-squares approach, does not exist. Under certain conditions, however, the two fundamental frequencies can be tracked in time by using estimates on each analysis frame as initial estimates in a refinement procedure for the next frame. In particular, if the analysis frames are closely spaced, then pitch changes slowly across two consecutive frames k and k +1. The pitch estimate obtained on frame k can then be used as the initial guess for estimating the pitch on frame (k + 1). A grid search is proposed as a means by which the tracking procedure be initialized. The iterative method of steepest descent[13] is then used for updating the pitch estimate on each frame. Figure 4-1 summarizes our approach to estimating pitch.

4.1 The Pitch Update Procedure

On each analysis frame, the method of steepest descent updates an initial pitch pair estimate by adding to the estimate a scaled error gradient with respect to the unknown pitch pair. Specifically, the pitch pair estimate on the kth frame is updated as

$$(\omega_a,\omega_b)_{k+1} = (\omega_a,\omega_b)_k - \alpha\Delta e \tag{4.1}$$

where Δe is the differential error. The error signal for the update (4.1) is the weighted least-mean square difference between the reconstructed waveform model estimate and the measured summed speech waveform, i.e., Equation (3.8), which is repeated here for convenience

$$e(n) = \min \sum_{n=-(N-1)/2}^{(N-1)/2} w(n)[x(n) - s(n)]^2 \tag{4.2}$$

The solution to x(n) in (4.2), $\hat{x}(n)$, has the form of (3.7), but with distinct frequencies replaced by multiples of a fundamental

$$\hat{x}(n) = \sum_{k=1}^{M_a} \hat{a}_k \cos(\omega_a kn + \hat{\phi}_{a,k}) + \sum_{k=1}^{M_b} \hat{b}_k \cos(\omega_b kn + \hat{\phi}_{b,k}) \tag{4.3}$$

and where, for a given pitch pair, the minimization in (4.2) takes place with respect to the unknown sine-wave amplitudes and phases. For a given pitch pair, the reconstructed waveform is obtained, therefore, by using the amplitudes and phases that result from the solution to the LSE problem, and thus the error surface over which we are minimizing is itself a minimum for each pitch pair.

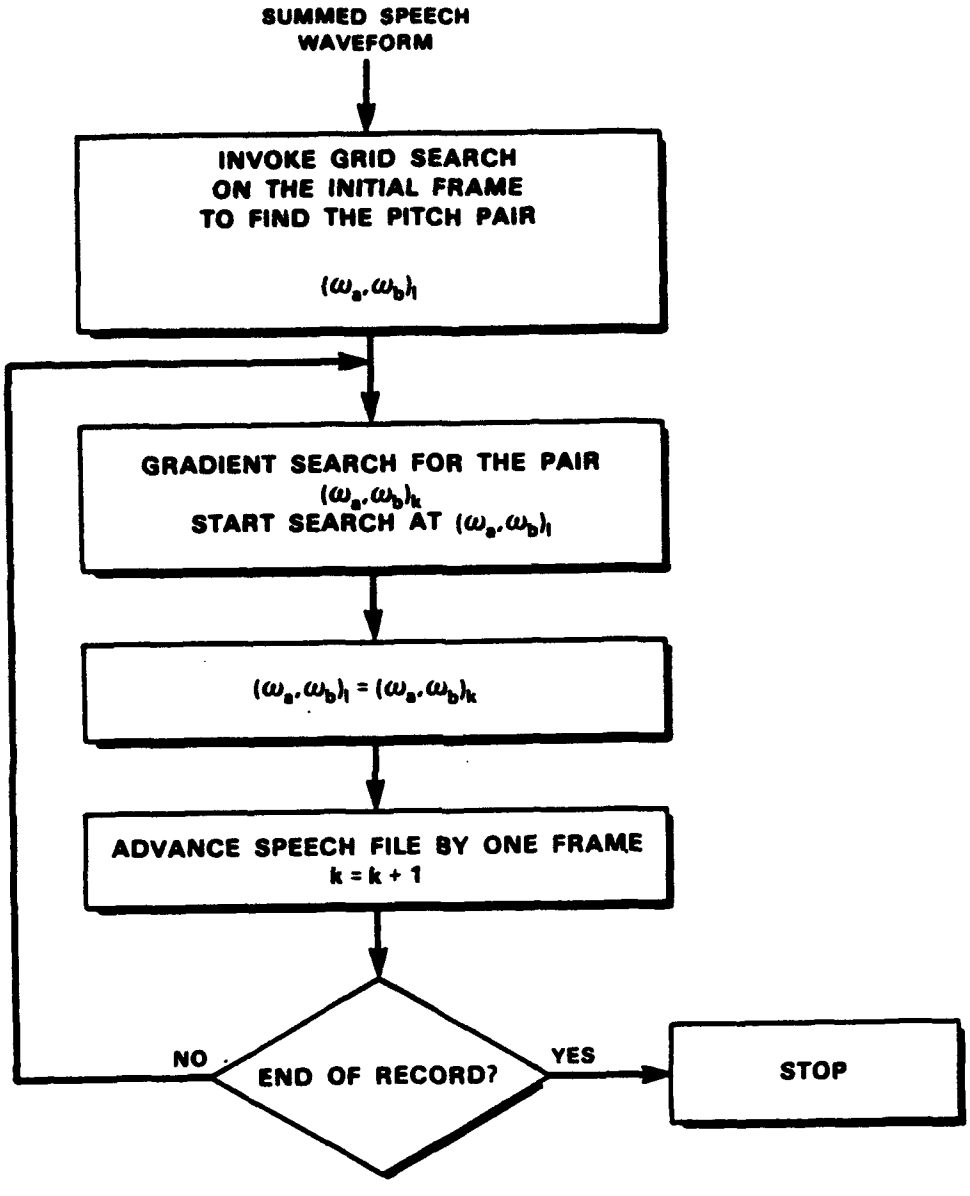


Figure 4-1. An algorithm for tracking the fundamental frequencies.

Figure 4-2 illustrates this iterative update procedure. The convergence factor, α, governs both the rate of convergence and the stability of the iteration. The gradient was approximated by the first central difference taken on a finely sampled error surface. Samples of the error surface, required by the central difference computation, were obtained by performing the LSE operation to determine the minimizing amplitudes and phases for the various pitch pairs shown in Figure 4-2. The iteration terminates once the error incurred on a step exceeds the error incurred on the preceding step. After each termination, the iteration can be restarted with a reduced step size Δ in order to obtain greater accuracy.

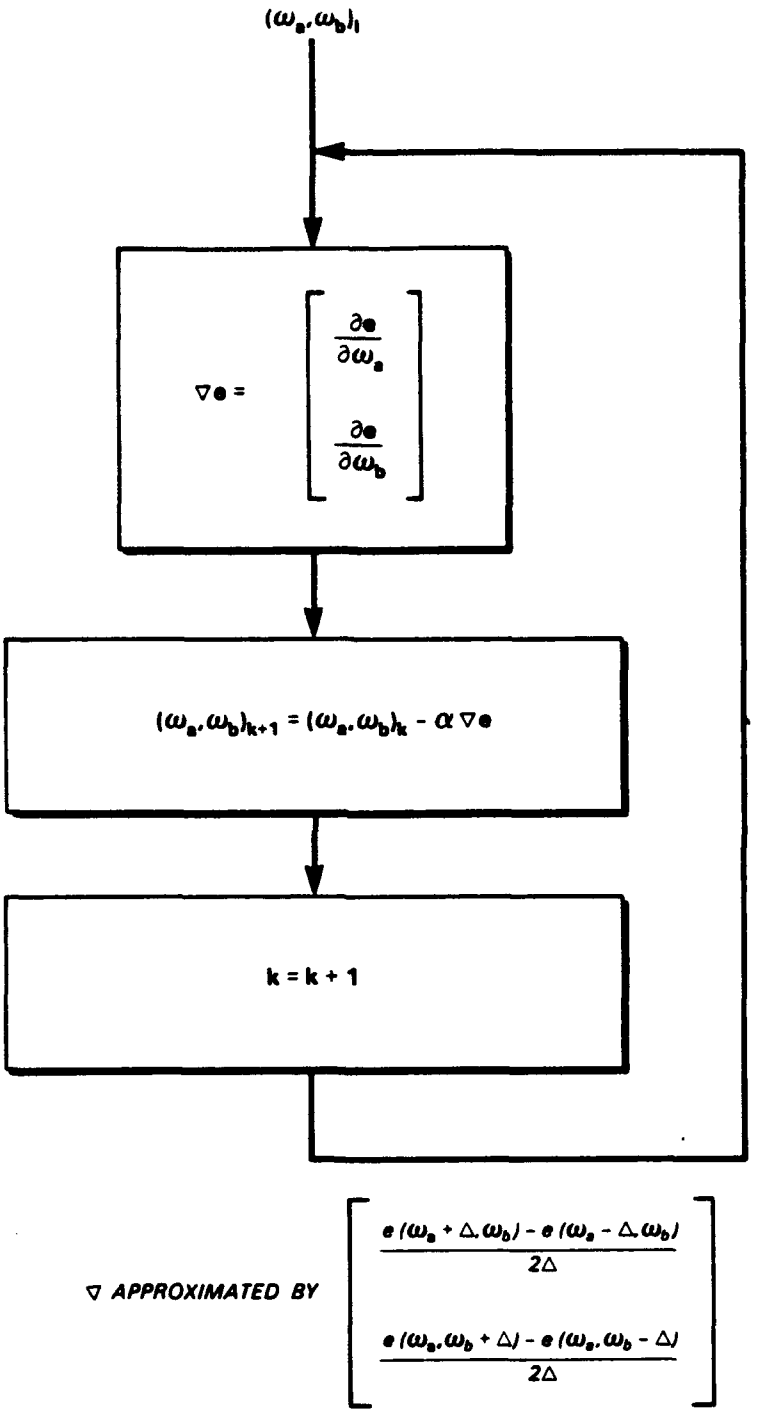
4.2 Estimation of an Initial Pitch Point

In order to initiate the pitch estimation algorithm, we must choose a pitch pair on the starting frame. Specifically, we sample the two-dimensional error surface in (4.2) and choose a pitch pair that yields the minimum error. Figure 4-3 illustrates a simple one-dimensional representation of the assumed error surface as a function of pitch. The function exhibits several extremum, but a single global minimum. A grid search must be based on a sampled version of this function. Figures 4-3(b) and 4-3(c) depict two different sampling intervals. In the first case, the sampling procedure indicates the region of convexity where the global minimum occurs. In the second instance, the search is too granular, and the minimum value of the sequence corresponds to a region of convexity that is attributed to a local, but not global, minimum. The severity of multiple local minimum is often tied to the high-frequency content in the signal.[14] The tracking procedure developed in the previous section will not tolerate this type of error. For this reason, a conservative sampling interval was chosen.[14]

A practical pitch range was given as a boundary to the grid search. For example, it might be known that speaker A's pitch is somewhere between 100 Hz and 150 Hz, and that speaker B's pitch is between 150 Hz and 220 Hz. This *a priori* information reduced the search range and circumvented the problem of examining candidate fundamentals which were factors of the underlying fundamentals.

4.3 An Example

For a limited data base, pitch contours were obtained by invoking the tracking procedure outlined in Sections 4.1 and 4.2. Figure 4-4 illustrates two pitch contours estimated from summed utterances of roughly equal intensities. The upper contour corresponds to the utterance, "We were away in Walla Walla," spoken by a female. The lower contour corresponds to the utterance, "Why were you away a year Roy?," spoken by a male. The solid line indicates the pitch that was extracted prior to summation.[14] The dotted line indicates the pitch that was extracted from the summation by means of the pitch tracking algorithm. The two initial points ($t = 0$) on the contours were extracted by means of the grid search described in Section 4.2.



89335-17

Figure 4-2. Gradient search. $e(\omega_1, \omega_2)$ is the error surface sampled at the pitch pair (ω_a, ω_b).

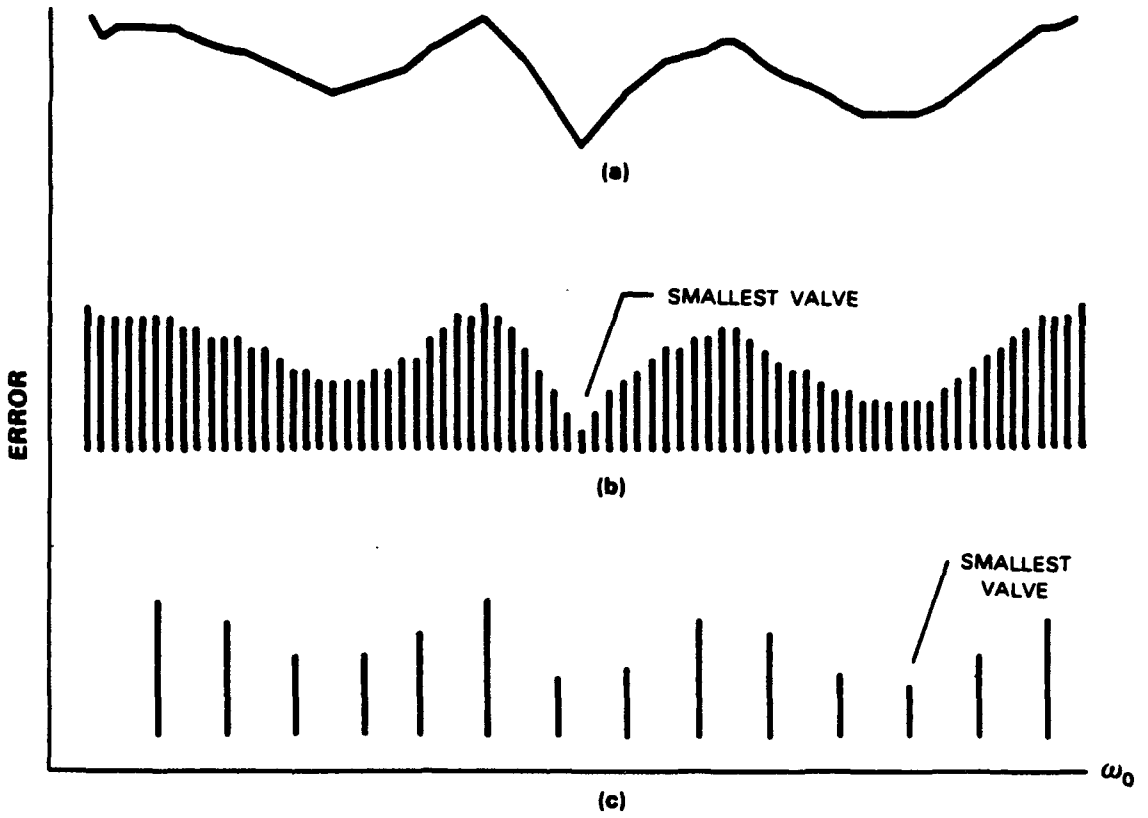


Figure 4-3. Sampling the error surface:(a) error surface, (b) fine sampling, and (c) sparse sampling.

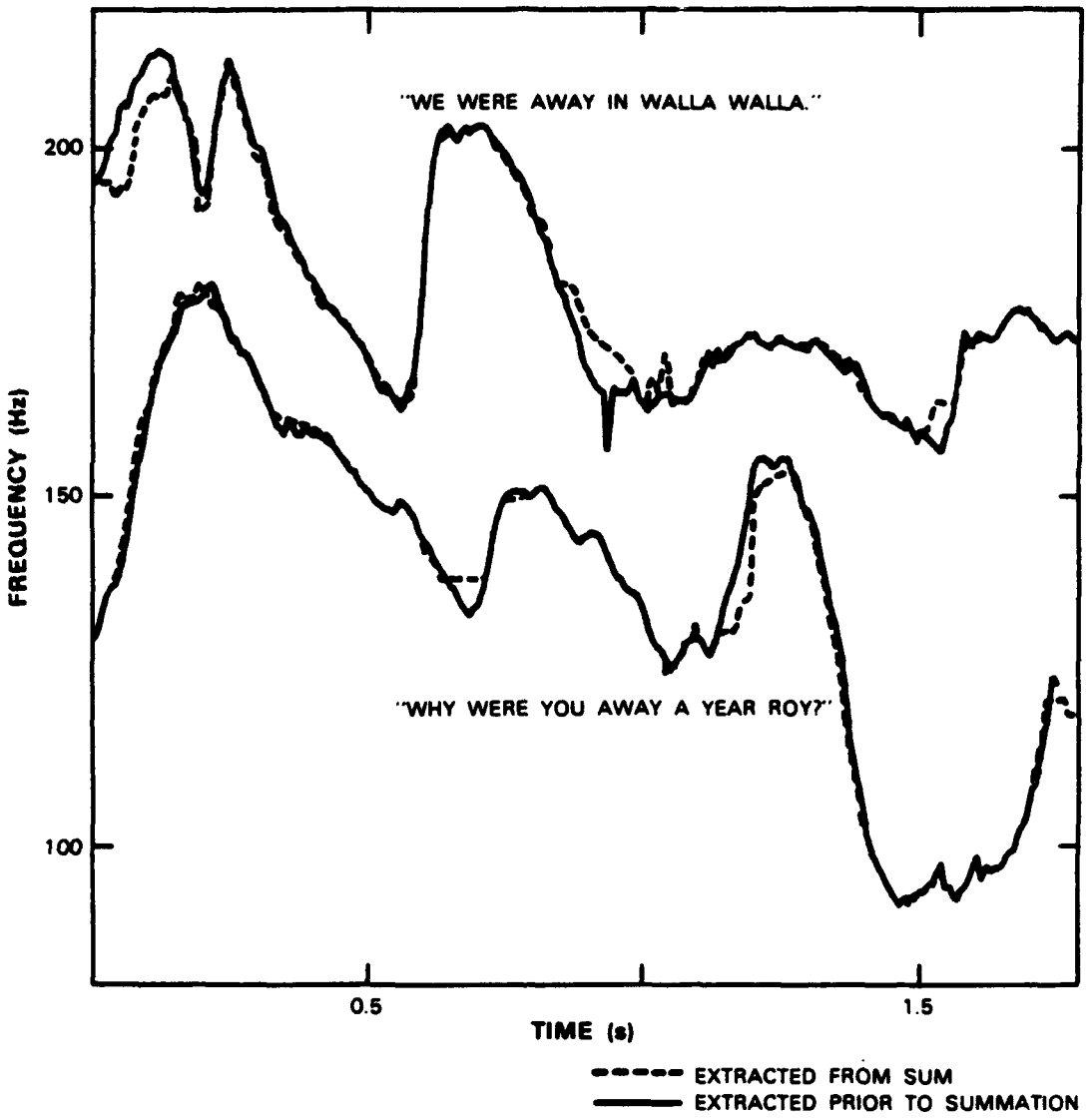


Figure 4-4. Two pitch contours extracted from summed vocalic waveforms.

4.4 Limitations

Although pitch extraction was successful on a number of all-voiced passages, the method suffers from the following disadvantages:

(1) The error from the LSE processor is used in pitch estimation; thus the pitch estimate is susceptible to matrix conditioning problems and lapses from stationarity where the periodic model breaks down. The data base for which pitch was extracted was limited to nonintersecting pitch contours. When pitch contours cross, the harmonic frequencies align, and the conditioning of the LSE problem deteriorates.

(2) The method of tracking the pitch contours from one frame to the next depends on smoothness and continuity, thus precluding pauses and consonants.

(3) The pitch algorithm was found useful for roughly a 3 dB difference in intensity between speakers. Larger intensity differences prohibited pitch tracking of the lower speaker.

5. MULTI-FRAME INTERPOLATION

We saw in Section 3.4 that the least-squares solution to estimating sine-wave amplitudes and phases can become ill-conditioned when sine-wave frequencies of the two underlying waveforms become arbitrarily close. In this section, we propose a synthesis scheme to help resolve the case where the ill-conditioning of the **H** matrix in (3.4) does not permit the solution to the LSE problem. This strategy exploits the time evolution of the sine-wave amplitudes and phases.

5.1 Approach

Assume the availability of either accurate frequency estimates or harmonic frequency estimates in the form of pitch contours. The LSE solution (3.2) is then used to extract the amplitudes and phases of a model with the given frequencies. Two conditions under which ill-conditioning of the **H** matrix can occur are illustrated in Figure 5-1.

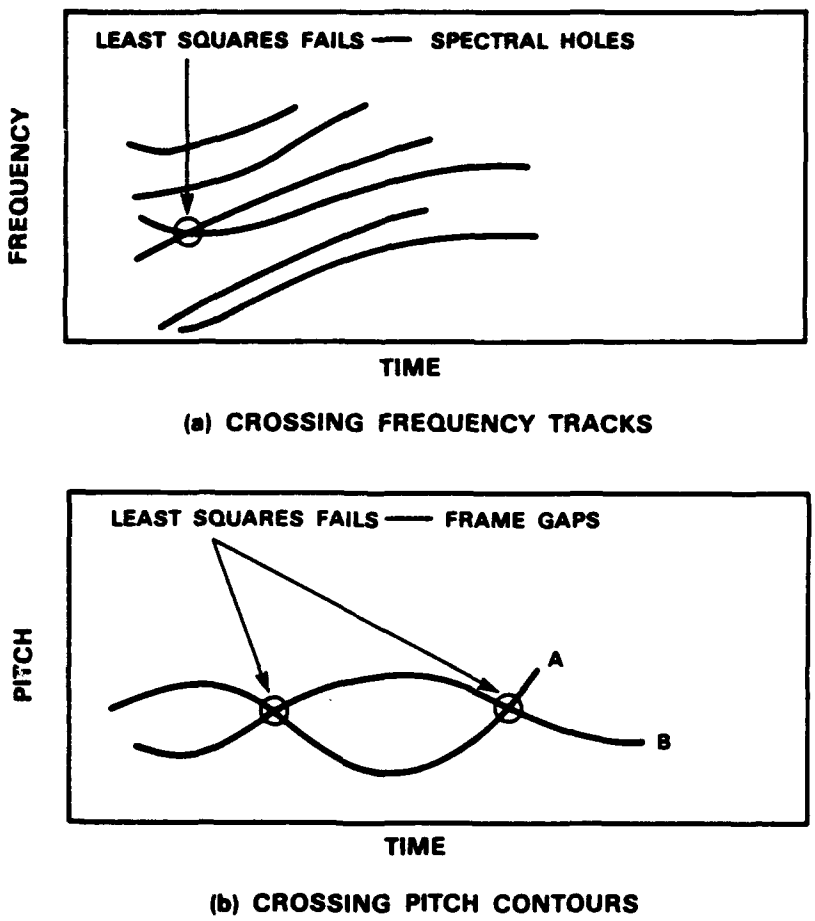


Figure 5-1. Failure of the least squares solution with closely-spaced frequencies.

In the first case, a phase and amplitude ambiguity occurs in solving (3.2) whenever isolated frequencies belonging to one speaker align themselves with frequencies belonging to the second speaker. In this case, the amplitude and phase separation cannot be resolved and, therefore, "spectral holes" arise. Ill conditioning may even occur when the two-pitch periods are markedly dissimilar. Consider the case of the two fundamental frequencies 100 Hz and 151 Hz. The two harmonic sets S_a and S_b can be generated.

$$S_a = (100, 200, 300, 400, 500, 600, 700) \quad (5.1a)$$

$$S_b = (151, 302, 453, 604, 755, 906, 1057) \quad (5.1b)$$

For this case, numerical instability arises at the overlapping pairs (300,302) and (600,604).

In the second case of Figure 5-1, ill conditioning occurs when the pitch contours cross. In this case all of the harmonics align. Therefore, separation is ambiguous at all frequencies and none of the sine-wave amplitudes and phases can be resolved. An entire frame of data is then deleted from the reconstruction.

The sinusoidal reconstruction strategy outlined in Section 2.1 can be used to interpolate component sine waves over regions of ill conditioning of the **H** matrix. If the kth frame contains frequencies that are too close to resolve, the corresponding amplitudes and phases are interpolated between frame (k – p) and frame (k + q), as in the reconstruction over a single frame (see Figure 5-2). Specifically, the linear amplitude and cubic phase interpolation strategies of Section 2.1 are used with the sine-wave amplitudes and phases, respectively, measured at the endpoints of frames (k – p) and (k + q). The integers, p and q, are chosen so that the frames (k – p) and (k + q) lie in regions where the amplitudes and phases can be resolved. This procedure is referred to as *multi-frame interpolation*. With multi-frame interpolation, the resulting frame interval is typically four frames or 20 ms, but can extend as long as 100 ms during stationary regions. A 25 Hz frequency "closeness" criterion was used in deciding when to perform interpolation.

When the pitch contours intersect, this procedure must be performed for every frequency component. When only a subset of the frequencies converge, the procedure is performed on only those frequencies. The algorithm was programmed to handle these different forms of interpolation at any one point in time. As illustrated in Figure 5-3, short or long interval interpolation, or no interpolation at all, can be performed along each frequency track. A long interpolation interval can occur in a steady-state region where two frequency tracks lie close to each other over a long period of time; while a short interval of interpolation can occur in rapidly varying regions where two frequency tracks cross at a sharp angle.

5.2 An Example

Figure 5-4(a) depicts a frame of vocalic speech (left) and the STFTM for that frame (right). Figure 5-4(b) depicts the reconstruction that is missing the fundamental frequency due to ill-conditioning of the **H** matrix. In Figure 5-4(c), the missing fundamental is resolved via multi-frame interpolation.

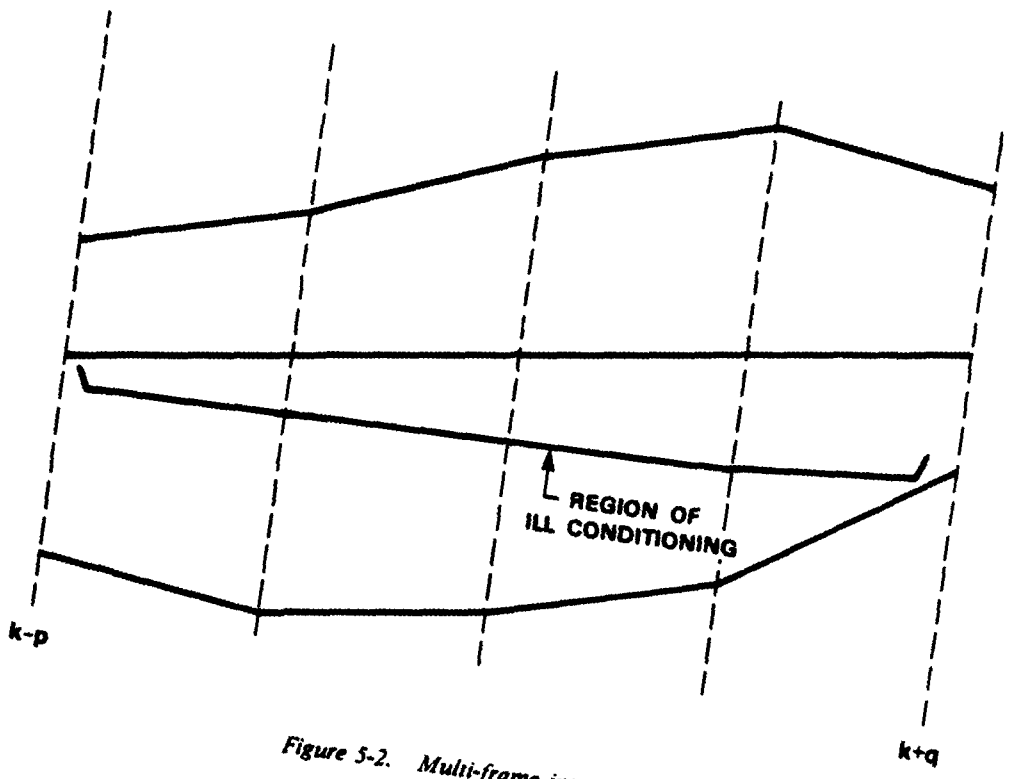


Figure 5-2. Multi-frame interpolation.

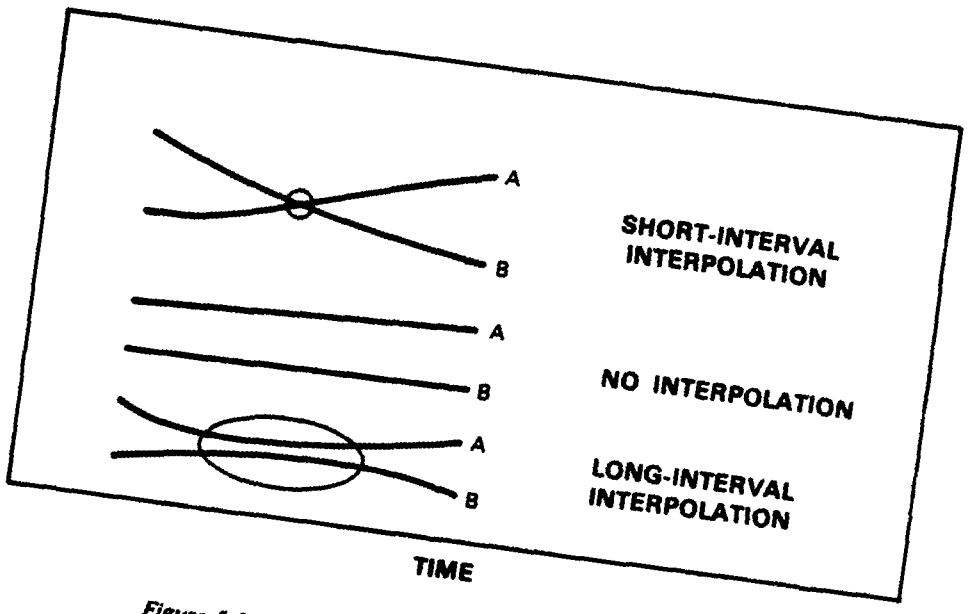


Figure 5-3. Different forms of multi-frame interpolation.

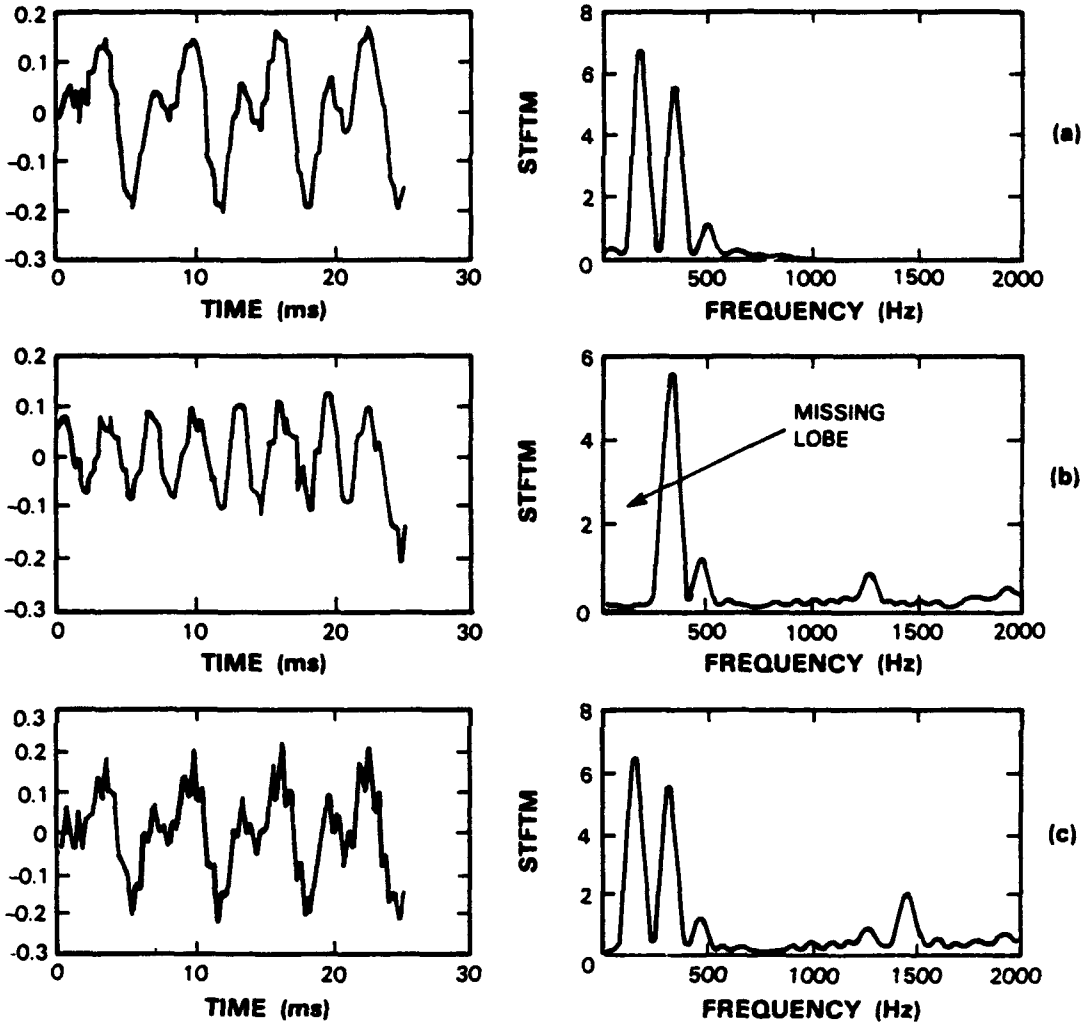


Figure 5-4. Recovery of missing lobe with multi-frame interpolation (MFI) (a) original, (b) no multi-frame interpolation, and (c) multi-frame interpolation.

6. EXPERIMENTAL RESULTS

This section summarizes the results of listening tests using the least squared error approach to sine-wave based speaker separation and enhancement. The LSE method is examined under the following three conditions: (1) *a priori* frequencies, (2) *a priori* pitch, and (3) estimated pitch. The importance of multi-frame interpolation is demonstrated.

6.1 Experimental Procedure

Microphone speech was digitized and sampled at 10 kHz. The data base was comprised of a variety of vocalic sentences; each sentence was recorded from male and female speakers. The average pitch of the speakers ranged between 100 and 200 Hz. Prior to summation, the sampled utterances were scaled to the desired target to interferer ratio. The calculation of this ratio was based on the long time average of the signal energy.

Four distinct sentences and three different speakers were used for the listening tests (see Table 6-1). Test utterances included the summed speech of male with male speakers, female with female speakers, and female with male speakers. The tests were conducted with ten listeners and consisted of two different forms. In one test (Test A), the listeners were given the original summed utterance followed by two enhanced versions (randomly arranged) of the target speaker. The listener was asked to choose the enhanced version he preferred in terms of intelligibility. In the second test (Test B), the listeners compared the intelligibility of the target speaker in the summed utterance to that in the processed version of the utterance and judged the improvement in the processed version on a scale of 0 to 3 given by (0) no improvement, (1) slight improvement, (2) definite improvement, and (3) significant improvement.

TABLE 6-1
Data Base Used in Listening Tests

A.	"We were away in Walla Walla." (male 1)
B.	"Our rule will lower your ear away." (female 1)
C.	"We were away in Walla Walla." (male 2)
D.	"All wear your ear low." (male 1)
E.	"Wear your ear low." (female 1)
F.	"All rare laws are well." (female 1)

The listening tests were performed within an acoustic sound chamber. The speech was played to the listener through a set of head phones. The listener was seated at a computer terminal which served as a means of cataloging his evaluation and also served as an interface to the sound system. The listener could monitor the summed utterance and the processed versions of the utterance at keyboard request, and as often as desired. The listener was prompted for a rating the moment he indicated, through the keyboard, that he was ready to evaluate the utterance. Once an utterance was evaluated, the listener could not return to that utterance.

Three of the four sentences were nonsensical to prohibit the listener from inferring the passage through context. The recurrence of the same sentences, however, allowed the listeners to learn the content of the passages as the test proceeded. Thus, ratings may not have been consistent throughout a single test.

6.2 Processing Schemes

The LSE processing scheme with *a priori* frequencies and *a priori* pitch, the LSE processing scheme with estimated pitch, and the frequency sampling scheme were evaluated. The LSE processing schemes were performed both with and without multi-frame interpolation.

In experiments with *a priori* frequencies, the LSE processor was provided with frequencies of all the major peaks of each speaker's STFTM prior to summation. The LSE processor extracted the corresponding amplitudes and phases of each utterance over successive frames. The frequencies, amplitudes, and phases associated with the target speaker were then given to the sinusoidal reconstruction system. In experiments with *a priori* pitch, the pitch contours were extracted by standard pitch estimation and by means of hand editing.[14] Integer multiples of the two fundamentals were used to parameterize the two frequency sets. The frequency sampling procedure described in Section 3.2 was invoked to determine whether the LSE approach yielded any improvement over a more direct strategy that made use of *a priori* frequencies, and also to provide a reference for all other processing schemes. In this method the STFT of the summed waveform was sampled at the STFTM peaks of the target speaker to obtain amplitude and phase samples that were given to the sine-wave synthesizer.

6.3 Listening Tests

The effect of processing was to significantly enhance the intelligibility of the target speaker, while reducing the interfering speech to "noise" or sometimes highly garbled cross talk. The remnant interfering speech was generally of much lower intensity than that perceived in the original sum, and in some cases, was not perceived as speech. Good signal recovery and intelligibility improvements were observed over a range of target-to-interferer ratios of 9 to -16 dB, although listening tests, described below, were performed at -16 dB.

Multi-Frame Interpolation

To determine the importance of multi-frame interpolation, summed utterances were formed with the target speaker 16 dB below the interferer. The test examined the effect of processing with *a priori* frequencies and *a priori* pitch, with and without multi-frame interpolation. Table 6-2 illustrates a strong preference for speech processed with multi-frame interpolation. Test A, in particular, showed that 81% of the listeners preferred multi-frame interpolation. Consequently, all following experiments were performed with the use of multi-frame interpolation.

TABLE 6-2
Tests Comparing Synthesis with Multi-Frame Interpolation (MFI) and No Multi-Frame Interpolation (NMFI)

TEST	MFI	NMFI	NO PREFERENCE
A	81%	6%	13%
B	2.12	1.6	*

Frequency Sampling

As a reference, speech was synthesized using sine-wave parameters derived from the frequency-sampling method. With the target speaker set 16 dB below the interfering speaker, the remnant of the interfering speaker was speech-like and of modest intensity. Only a minimal degree of enhancement is obtained; an average rating level of .6 was achieved using Test B.

A Priori Frequencies and Pitch

In the next set of experiments, the result of LSE processing with *a priori* frequencies was compared with processing with *a priori* pitch, again with a 16 dB intensity difference. Test B shows significant intelligibility gains in both cases. As illustrated in Table 6-3, the pitch-based system, however, was not capable of attaining the same level of enhancement as the frequency-based system. Additional cross talk was present due to inaccuracy in the pitch estimate and limitation to a harmonic frequency set. Nevertheless, significant intelligibility improvements were attained.

Estimated Pitch

Since the pitch extraction algorithm has been successfully demonstrated at only roughly equal intensity levels, a reference was first formed by processing with *a priori* frequencies and

TABLE 6-3
Results of Listening Tests Comparing Synthesis with *A Priori* Frequencies and *A Priori* Pitch

TEST	FREQUENCY	PITCH	NO PREFERENCE
A	72%	0%	28%
B	2.6	1.64	*

a priori pitch at equal intensity settings. The LSE processor achieved significant suppression of the interferer with both *a priori* frequency and pitch. Since this case does not occur often in practice, extensive listening tests were not performed. The same set of utterances was then processed where *a priori* information consisted of only an initial point on the two dimensional pitch contour. This initial point was obtained through the use of the grid search of Section 4.2. A pair of pitch contours was then generated from the summed speech waveform by the method described in Section 4.1. Significant suppression was obtained in the reconstruction of the desired speaker; although, in contrast to the case where *a priori* pitch was assumed, some quality loss was apparent.

The system was capable of handling only utterances having pitch contours that followed the conditions given in Section 4. The system was not capable of resolving situations in which the two pitch contours crossed. The experiments were performed on summed male and female speech, since, in such cases, the crossing of pitch contours is less likely to occur.

6.4 Assessment

There are three principal reasons for the cross talk that remains after processing: modeling errors, frequency errors, and unresolvable parameters.

(1) Modeling errors occur when the steady-state sine-wave model is not capable of accurately representing an entire frame of speech, as when the vocal tract or excitation changes too rapidly.

(2) Degradation arises when the model is not parameterized with the correct frequencies. This problem was made apparent when the reconstruction with *a priori* frequencies was compared with reconstruction with *a priori* pitch and estimated pitch.

(3) The multi-frame interpolation strategy helps to resolve cases with closely spaced frequencies, but if the effective frame length becomes too large, degradation may occur.

7. DISCUSSION

This report described a method for talker interference suppression based on a sinusoidal representation of speech. A least squares approach for obtaining the sine-wave parameters was proposed. When sine-wave frequencies of the underlying waveforms were closely spaced, a multi-frame interpolation scheme was used to recover missing spectral regions which could not be resolved by the LSE method.

Although success of the LSE relies on accurate frequency estimates, results of this report indicate that pitch estimates can be used in place of frequency estimates. Pitch extraction is also necessary as a means of partitioning the frequency set into a subset that is attributed to speaker A and a subset that is attributed to speaker B. Thus, the further development of a pitch estimation algorithm, capable of handling summed waveforms of vastly different intensity levels, is critical to sinusoidal speech separation and enhancement.

The results of this report lead to a number of other important areas of continued research. The experiments at the onset of this report using the peak-picking and frequency-sampling methods raise some fundamental questions about the methods themselves, about the limitations of sine-wave based analysis-synthesis, and about how more accurate parameter estimation may be achieved in this context. These methods, as well as the LSE approach, relied on local characteristics of the STFT, i.e., separation was attempted without recourse to correlation in the STFT across frequency. A drawback to this approach was manifested in solving for the sine-wave parameters by the LSE method in the presence of closely spaced frequencies. One form of accounting for spectral correlation might take the form of estimating amplitude and phase *envelopes* of the STFT of each speech waveform, along with sine-wave frequencies or pitch. Such estimation might use envelope models or template-based envelope matching. It is also of interest to develop techniques for joint estimation of sinusoidal amplitudes, frequencies, and phases which include multi-frame continuity constraints and interpolation strategies.

This report was limited to utterances for which a vocalic excitation was present. There are a host of utterances for which such an excitation is not present. Examples include fricatives, stop consonants, and whispers. Separation of such combined utterances will require relaxing the harmonic model, using more complex modeling of the speech waveform, and jointly estimating sinusoidal model parameters and voicing states of target and interfering signals.

APPENDIX A

In this Appendix, we demonstrate the relation between the various sine-wave based representations of x(n) used in Sections 2 and 3. From a standard trigonometric identity, we can write (2.2)

$$x(n) = \sum_{k=1}^{M} a_k \cos(\omega_k n + \phi_k)$$

$$= \sum_{k=1}^{M} a_k \cos(\phi_k) \cos(\omega_k n)$$

$$= \sum_{k=1}^{M} a_k \sin(\phi_k) \sin(\omega_k n)$$

$$= \sum_{k=0}^{M} \alpha_k \cos(\omega_k n) + \sum_{k=0}^{M} \beta_k \sin(\omega_k n) \qquad \text{(A1.a)}$$

where

$$\alpha_k = a_k \cos(\phi_k)$$

$$\beta_k = -a_k \sin(\phi_k) \qquad \text{(A1.b)}$$

Alternatively, the sequence x(n) can be written

$$x(n) = \sum_{k=1}^{M} \frac{a_k}{2} e^{j\phi_k} e^{j\omega_k n} + \sum_{k=1}^{M} \frac{a_k}{2} e^{-j\phi_k} e^{-j\omega_k n} \qquad \text{(A2)}$$

Assuming negligible contribution from the negative frequency terms in (A2), then the STFT of x(n), $S(\omega)$, for $\omega = \omega_k > 0$ can be written as

$$\sum_{\ell=1}^{M} \frac{a_\ell}{2} e^{j\phi_\ell} W(\omega_k - \omega_\ell) = S(\omega_k) \qquad \text{(A3.a)}$$

or

$$\sum_{\ell=1}^{M} \frac{a_\ell}{2} \cos(\phi_\ell) W(\omega_k - \omega_\ell) = \mathrm{Re}[S(\omega_k)]$$

$$\sum_{\ell=1}^{M} \frac{a_\ell}{2} \sin(\phi_\ell) W(\omega_k - \omega_\ell) = \mathrm{Im}[S(\omega_k)] \qquad \text{(A3.b)}$$

and therefore,

$$\sum_{\ell=1}^{M} \alpha_\ell \; W(\omega_k - \omega_\ell) = 2\mathrm{Re}\,[S(\omega_k)]$$

$$\sum_{\ell=1}^{M} \beta_\ell \; W(\omega_k - \omega_\ell) = -2\mathrm{Re}\,[S(\omega_k)] \qquad \text{(A3.c)}$$

where,

$$\alpha_\ell = a_\ell \cos(\phi_\ell)$$

$$\beta_\ell = -a_\ell \sin(\phi_\ell) \qquad \text{(A3.d)}$$

APPENDIX B
THE LEAST SQUARED ERROR SOLUTION

The steady state model for a frame of summed vocalic speech is expressed as a sum of sinusoids.

$$x(n) = \sum_{k=1}^{M} c_k \cos(\omega_k n + \phi_k) \tag{B.1}$$

Alternately, (B.1) may be written in terms of quadrature components.

$$x(n) = \sum_{k=0}^{M} \alpha_k \cos(\omega_k n) + \sum_{k=0}^{M} \beta_k \sin(\omega_k n) \tag{B.2}$$

$$c_k = \alpha_k^2 + \beta_k^2 \qquad \phi_k = \tan^{-1}(-\beta/\alpha)$$

In this form, it is easy to see that x(n) is a linear function of the unknown coefficients α_k and β_k. Let N be the length of the frame. Assume that N is odd. This appendix will determine the coefficients α_k and β_k that yield the best fit of the model, (B.2), to a single frame of additive vocalic speech, over the region $-(N-1)/2 < n \leqslant (N-1)/2$, when the frequencies ω_k are given.

Before proceeding, we introduce some vector notation that will allow concise statement of the linear least squares problem. Let the vectors $\underline{s}$ and $\underline{x}$ be the sampled waveform and the sampled model. The set of sinusoidal frequencies will be denoted by the vector $\underline{\omega}$.

$$\underline{s} = \begin{bmatrix} s[-(N-1)/2] \\ s[1-(N-1)/2] \\ s[2-(N-1)/2] \\ \vdots \\ s[n-(N-1)/2] \\ \vdots \\ s[(N-1)/2] \end{bmatrix} \quad \underline{x} = \begin{bmatrix} x[-(N-1)/2] \\ x[1-(N-1)/2] \\ x[2-(N-1)/2] \\ \vdots \\ x[n-(N-1)/2] \\ \vdots \\ x[(N-1)/2] \end{bmatrix} \quad \underline{\omega} = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ \vdots \\ \omega_n \\ \vdots \\ \omega_M \end{bmatrix} \tag{B.3}$$

Let the vectors $\underline{r}(\omega_k)$ and $\underline{d}(\omega_k)$ denote the respective time sequences of cosine and sine functions.

$$\underline{r}(\omega) = \begin{bmatrix} \cos\left[\omega \dfrac{-(N-1)}{2}\right] \\ \cos\left(\omega\left[1 - \dfrac{(N-1)}{2}\right]\right) \\ \cos\left(\omega\left[2 - \dfrac{(N-1)}{2}\right]\right) \\ \vdots \\ \cos\left[\omega \dfrac{(N-1)}{2}\right] \end{bmatrix} \quad \underline{d}(\omega) = \begin{bmatrix} \sin\left[\omega \dfrac{-(N-1)}{2}\right] \\ \sin\left(\omega\left[1 - \dfrac{(N-1)}{2}\right]\right) \\ \sin\left(\omega\left[2 - \dfrac{(N-1)}{2}\right]\right) \\ \vdots \\ \sin\left[\omega \dfrac{(N-1)}{2}\right] \end{bmatrix} \tag{B.4}$$

The vectors $\underline{\alpha}$ and $\underline{\beta}$ will be the coefficient vectors for the cosine and sine terms in the model (B.2). The matrices $R(\underline{\omega})$ and $D(\underline{\omega})$ will have $\underline{r}(\omega_k)$ and $\underline{d}(\omega_k)$ for their respective kth columns.

$$R(\underline{\omega}) = \left[[\underline{r}(\omega_1) \mid \underline{r}(\omega_2) \mid \underline{r}(\omega_3) \mid \ldots \underline{r}(\omega_M)] \right] \tag{B.5}$$

$$[R(\underline{\omega})]_{n,k} = \cos\left[\omega_k \left(n - \frac{N+1}{2}\right)\right] \tag{B.6}$$

$$D(\underline{\omega}) = \left[\underline{d}(\omega_1) \mid \underline{d}(\omega_2) \mid \underline{d}(\omega_3) \mid \ldots \underline{d}(\omega_M) \right] \tag{B.7}$$

$$[D(\underline{\omega})]_{n,k} = \sin\left[\omega_k \left(n - \frac{N+1}{2}\right)\right] \tag{B.8}$$

The notation $R(\underline{\omega})$ suggests the dependency of the matrix on $\underline{\omega}$. This parenthetical notation will be dropped for convenience, since the vector of frequencies is considered to be a fixed parameter throughout.

The model (B.2) can now be concisely expressed in terms of the adopted vector notation.

$$\underline{x} = R\underline{\alpha} + D\underline{\beta} \tag{B.9}$$

The general least squares problem can be stated as follows:

$$\underline{v} = \min_{\underline{v}} [\underline{x}(\underline{v}) - \underline{s}]^T \, W[\underline{x}(\underline{v}) - \underline{s}] \tag{B.10}$$

where $\underline{x}(\underline{v})$ is the model parameterized by the vector $\underline{v}$ consisting of the vectors $\underline{\alpha}$ and $\underline{\beta}$. The matrix W is a positive definite diagonal weighting matrix with diagonal terms equal to the analysis window.

$$W = \begin{bmatrix} w[-(N-1)/2] & 0 & 0 & & \\ 0 & w[1-(N-1)/2] & 0 & & \\ 0 & 0 & w[2-(N-1)/2] & & \\ 0 & 0 & 0 & \cdot & \\ & \cdot & & \cdot & \\ & \cdot & & & \\ & \cdot & & & w[(N-1)/2] \end{bmatrix} \tag{B.11}$$

Since the model is a linear function of the parameter vector, the model can be written as

$$\underline{x} = F\underline{v} \tag{B.12}$$

where F is given as a partitioned matrix.

$$F = [R \mid D] \tag{B.13}$$

The solution of the linear least squares problem (B.10) is well known, and given by the solution to the following matrix equation.

$$F^T W F \underline{v} = F^T W \underline{s} \tag{B.14}$$

so that $\underline{v}$ can be written in terms of the inverse of $F^T WF$. This inverse will exist if the columns of F are linearly independent.[13]

$$\underline{v} = [F^T WF]^{-1} F^T W \underline{s} \tag{B.15}$$

Because of the orthogonality between the sampled cosine and sine terms, and because the weighting is symmetric about the origin, the partitioned matrix of inner products will have no cross terms.

$$F^T WF = \begin{bmatrix} R^T WR & 0 \\ 0 & D^T WD \end{bmatrix} \tag{B.16}$$

Thus, the LSE solution can be written in terms of two independent expressions. The first equation will yield the coefficient vector $\underline{\alpha}$. The second equation can be solved for the coefficient vector of the quadrature terms $\underline{\beta}$.

$$\underline{\alpha} = [R^T WR]^{-1} R^T W \underline{s} \tag{B.17}$$

$$\underline{\beta} = [D^T WD]^{-1} D^T W \underline{s} \tag{B.18}$$

In order to interpret this result in the context of Section 3, it will be helpful to rewrite the short time Fourier transform (STFT) in terms of its real and imaginary parts.

$$S_{re}(\omega) = \sum_{n=-(N-1)/2}^{(N-1)/2} s(n)\, w(n) \cos(\omega n) \tag{B.19}$$

$$S_{im}(\omega) = -\sum_{n=-(N-1)/2}^{(N-1)/2} s(n)\, w(n) \sin(\omega n) \tag{B.20}$$

This can be expressed in terms of the adopted vector notation.

$$S_{re}(\omega) = \underline{r}(\omega)^T W \underline{s} \tag{B.21}$$

$$S_{im}(\omega) = -\underline{d}(\omega)^T W \underline{s} \tag{B.22}$$

The real and imaginary parts of the STFT sampled at the frequencies $\omega_1, \omega_2, \omega_3, \ldots \omega_M$, can be expressed as a vector function of a vector variable.

$$\underline{S}_{re}(\underline{\omega}) = \begin{bmatrix} S_{re}(\omega_1) \\ S_{re}(\omega_2) \\ S_{re}(\omega_3) \\ \vdots \\ S_{re}(\omega_k) \\ \vdots \\ S_{re}(\omega_{M-1}) \end{bmatrix} = R^T W \underline{s} \tag{B.23}$$

and likewise for the imaginary component. From Equation (B.17) and Equation (B.23), it should be clear that the vector $\underline{\alpha}$ is obtained by pre-multiplying a vector which contains frequency samples of the real part of the STFT by the matrix $[R^T W R]^{-1}$. The parameter vector $\underline{\beta}$ is similarly obtained by pre-multiplying a vector which contains frequency samples of the imaginary part of the STFT by the matrix $-[D^T W D]^{-1}$.

It is left to show that

$$R^T W R \approx D^T W D \approx \frac{1}{2} H \quad , \tag{B.24}$$

for the window lengths, frequency spacing, and window transforms considered in Section 3. In Section 3, H was a matrix that contained the arrangement of window transform samples.

$$H = \begin{bmatrix} W(0) & W(\omega_1 - \omega_2) & W(\omega_1 - \omega_3) & \cdots & W(\omega_1 - \omega_M) \\ W(\omega_2 - \omega_1) & W(0) & W(\omega_2 - \omega_3) & & W(\omega_2 - \omega_M) \\ W(\omega_3 - \omega_1) & W(\omega_3 - \omega_2) & W(0) & & W(\omega_3 - \omega_M) \\ & & \vdots & & \\ W(\omega_M - \omega_1) & W(\omega_M - \omega_2) & W(\omega_M - \omega_3) & & W(0) \end{bmatrix} \tag{B.25}$$

Now since

$$[WR]_{n,k} = w[n - (N + 1)/2] \cos \left[\omega_k (n - \frac{N + 1}{2})\right] \tag{B.26}$$

then

$$[R^T WR]_{\ell,k} = \sum_{n=-(N-1)/2}^{(N-1)/2} \cos(\omega_\ell n)\, w(n) \cos(\omega_k n) \tag{B.27}$$

It is apparent that (B.27) is the cosine transform of the windowed cosine sequence, $w(n) \cos(\omega_k n)$. Therefore, it must equal the real part of the Fourier transform of that sequence. Using the modulation property of the Fourier transform, along with the fact that the window transform is purely real.

$$[R^T WR]_{\ell,k} = \frac{1}{2} W(\omega_\ell - \omega_k) + \frac{1}{2} W(\omega_\ell + \omega_k) \quad . \tag{B.28}$$

Equation (B.28) gives the entries in the matrix for the exact solution to the LSE problem. The approximation used in Section 3 neglected the term $1/2\, W(\omega_\ell + \omega_k)$; that is,

$$[R^T WR]_{\ell,k} \approx \frac{1}{2} W(\omega_\ell - \omega_k) \quad . \tag{B.29}$$

This approximation is valid when the window length is sufficient (i.e., the bandwidth of the window transform is sufficiently narrow). The approximation was primarily introduced for reasons of style. It allows for a less cluttered, more intuitive argument in Section 3. The exact matrix entries (B.28) were used in all algorithm simulations.

APPENDIX C
A SENSITIVITY CALCULATION

In this appendix, it is shown that ill-conditioning of the least squares error solution increases as the spacing between the sine-wave frequencies decreases. Consider the solution to separation of two overlapping lobes as described by Figure 3-1 with $\omega_1 = \omega$ and $\omega_2 = \omega + \Delta$.

$$\begin{bmatrix} 1 & W(\Delta) \\ W(\Delta) & 1 \end{bmatrix} \begin{bmatrix} S_a(\omega) \\ S_b(\omega + \Delta) \end{bmatrix} = \begin{bmatrix} S(\omega) \\ S(\omega + \Delta) \end{bmatrix} \tag{C.1}$$

The solution to this matrix equation is obtained by inverting the matrix that appears on the left of (C.1). Let this matrix be denoted by $\mathbf{H}(\Delta)$ to make explicit the dependence of the matrix upon the frequency difference. The solution to (C.1) is given by,

$$\begin{bmatrix} S_a(\omega) \\ S_b(\omega + \Delta) \end{bmatrix} = H^{-1}(\Delta) \begin{bmatrix} S(\omega) \\ S(\omega + \Delta) \end{bmatrix} . \tag{C.2}$$

The partial derivative of the right side of this matrix equation, with respect to Δ, can be expressed as follows:

$$\frac{\partial}{\partial \Delta} \begin{bmatrix} S_a(\omega) \\ S_b(\omega + \Delta) \end{bmatrix} = -H^{-1}(\Delta) \left[\frac{\partial}{\partial \Delta} H(\Delta) \right] H^{-1}(\Delta) \begin{bmatrix} S(\omega) \\ S(\omega + \Delta) \end{bmatrix} + H(\Delta)^{-1} \begin{bmatrix} 0 \\ \frac{\partial}{\partial \Delta} S(\omega + \Delta) \end{bmatrix} \tag{C.3}$$

which can be expanded as,

$$\frac{-1}{|\det H(\Delta)|^2} \begin{bmatrix} 1 & -W(\Delta) \\ -W(\Delta) & 1 \end{bmatrix} \begin{bmatrix} 0 & \frac{\partial}{\partial \Delta} W(\Delta) \\ \frac{\partial}{\partial \Delta} W(\Delta) & 0 \end{bmatrix} \begin{bmatrix} 1 & -W(\Delta) \\ -W(\Delta) & 1 \end{bmatrix} \begin{bmatrix} S(\omega) \\ S(\omega + \Delta) \end{bmatrix} \tag{C.4}$$

$$+ \frac{1}{|\det H(\Delta)|} \begin{bmatrix} 1 & -W(\Delta) \\ -W(\Delta) & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \frac{\partial}{\partial \Delta} S(\omega + \Delta) \end{bmatrix}$$

where,

$$\det H(\Delta) = 1 - W^2(\Delta).$$

When $\Delta \to 0$, $\det H(\Delta) \to 0$. Therefore, as the frequency spacing tends to zero, the data independent factor in the second term in Equation (C.4) tends to infinity. For small Δ, the matrix equation is poorly conditioned; the solution is sensitive to small perturbations in Δ.

ACKNOWLEDGMENTS

The authors wish to thank Bob McAulay and Cliff Weinstein for helpful discussions.

REFERENCES

1. B.A. Hanson and D.Y. Wong, "Processing Techniques for Intelligibility Improvement to Speech with Co-Channel Interference," Final Technical Report, Rome Air Development Center (September 1983), DTIC AD-A135702.

2. R.J. McAulay and T.F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," Technical Report 693, Lincoln Laboratory, MIT (17 May 1985) and in *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-34, No. 4, August 1986, DTIC AD-A157023.

3. T.F. Quatieri and R.J. McAulay "Speech Transformations Based on a Sinusoidal Representation, Technical Report 717, Lincoln Laboratory, MIT (16 May 1986) and in *IEEE Trans. Acoust. Speech Signal Process,* Vol. ASSP-34, No. 6, December 1986, DTIC AD-A169740.

4. R.H. Frazier, S. Samsam, L.D. Braida, A.V. Oppenheim, "Enhancement of Speech by Adaptive Filtering," *Proceedings of IEEE Intl. Conf. on Acoust. Speech and Signal Process.* (IEEE, New York, 1980), pp. 251-253.

5. V.C. Shields, "Separation of Additive Speech Signals by Digital Comb Filtering," S.M. Thesis, Department of Electrical Engineering and Computer Science, MIT (September 1970).

6. T.W Parsons and M.R. Weiss, "Enhancing/Intelligibility of Speech in Noisy or Multi-Talker Environments," Final Technical Report, Rome Air Development Center (June 1975), DTIC AD-A013767.

7. T.W Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection," J. Acoust. Soc. Am. **60**, 911 (1976).

8. J. Naylor and S.F. Boll, "Simultaneous Talker Separation, "Final Program and Paper Summaries for the 1986 Digital Signal Processing Workshop, Chatham, MA, pp. 5.7.1.-5.7.2.

9. J. Naylor and S.F. Boll, "Techniques for Suppression of an Interfering Talker in Co-Channel Speech," Intl. Conf. on Acoust. Speech and Signal Process., Dallas, Texas, April 1983, Vol. 1, pp. 205-208.

10. D.G. Childers and C.K. Lee, "Co-Channel Speech Separation," Intl. Conf. on Acoust. Speech and Signal Process., Dallas, Texas, April 1987, Vol. 1, pp. 181-184.

11. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, New Jersey, 1978).

12. A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, New Jersey, 1975).

13. G.S. Strang, *Linear Algebra and Its Applications* (Academic Press, 1980).

14. R.G. Danisewicz, "Speaker Separation of Steady State Vowels," S.M. Thesis, Department of Electrical Engineering and Computer Science, MIT, June 1987.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Report 794	5. MONITORING ORGANIZATION REPORT NUMBER(S) ESD-TR-87-095
6a. NAME OF PERFORMING ORGANIZATION Lincoln Laboratory, MIT 6b. OFFICE SYMBOL *(If applicable)*	7a. NAME OF MONITORING ORGANIZATION Electronic Systems Division
6c. ADDRESS *(City, State, and Zip Code)* P.O. Box 73 Lexington, MA 02173-0073	7b. ADDRESS *(City, State, and Zip Code)* Hanscom AFB, MA 01731
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Systems Command, USAF 8b. OFFICE SYMBOL *(If applicable)*	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS *(City, State, and Zip Code)* Andrews AFB Washington, DC 20334	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: 33401F, 64754F PROJECT NO.: 51 TASK NO.: WORK UNIT ACCESSION NO.:

11. TITLE *(Include Security Classification)*

An Approach to Co-Channel Talker Interference Suppression Using a Sinusoidal Model for Speech

12. PERSONAL AUTHOR(S)

Danisewicz, Ronald G. and Quatieri, Thomas F.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT *(Year, Month, Day)*	15. PAGE COUNT
Technical Report	FROM _____ TO _____	5 February 1988	64

16. SUPPLEMENTARY NOTATION

None

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

Co-Channel Talker Interference Suppression; Speaker Separation; Sinusoidal Model; Sinusoidal Analysis-Synthesis; Minimum Mean-Squared Error Estimation; Least-Squares Estimation; Speech Enhancement; Time Evolution; Multi-Frame Interpolation; Closely-Spaced Frequencies; Pitch Estimation; Pitch Tracking; Summed Vocalic Speech Waveforms

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This report describes a new approach to co-channel talker interference suppression based on a sinusoidal representation of speech, which has been applied effectively in situations where both the desired and interfering speech waveforms are vocalic. The technique fits a sinusoidal model to additive vocalic speech segments such that the least mean squared error between the model and the combined waveforms is obtained. Enhancement is achieved by synthesizing a waveform from the sine waves attributed to the desired speaker. Least squares estimation is applied to obtain sine-wave amplitudes and phases of both talkers, based on either *a priori* sine-wave frequencies or *a priori* fundamental frequency contours. When the frequencies of the two waveforms are closely spaced, the least squares approach can have difficulty in tracking the sine-wave parameters. In these cases, the performance is significantly improved by an interpolation technique which predicts the time evolution of the sinusoidal parameters across multiple analysis frames. The approach yielded good suppression of the interfering speech and enhancement of the target speech over a wide range (9 to -16 dB) of target-to-interferer ratios. The least-squared error approach is also extended to estimate fundamental frequency contours of both speakers from the summed waveform, and applied further to estimate the remaining sinusoidal parameters.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt. Col. Hugh L. Southall, USAF	22b. TELEPHONE *(Include Area Code)* (617) 981-2330 22c. OFFICE SYMBOL ESD/TML